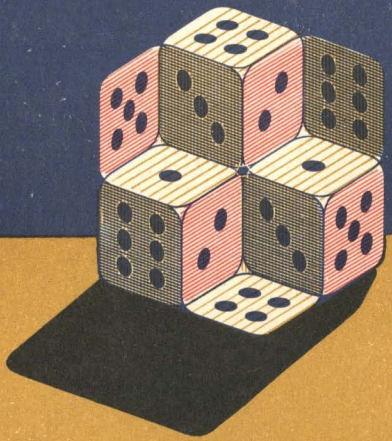
Probability Theory

(first steps)

by E. S. Wentzel



Mir Publishers · Moscow


ИЗДАТЕЛЬСТВО
МИР

Е. С. Вентцель

Теория вероятностей

(Первые шаги)

Издательство «Знание»
Москва

Probability Theory

(first steps)

by E.S.Wentzel

Translated from the Russian
by
N. Deineko

Mir Publishers · Moscow

First published 1982
Revised from the 1977 Russian edition
Second printing 1986

На английском языке

Contents

Probability Theory and Its Problems

Probability theory occupies a special place in the family of mathematical sciences. It studies special laws governing random phenomena. If probability theory is studied in detail, then a complete understanding of these laws can be achieved. The goal of this book is much more modest—to introduce the reader to the basic concepts of probability theory, its problems and methods, possibilities and limitations.

The modern period of the development of science is characterized by the wide application of probabilistic (statistical) methods in all branches and in all fields of knowledge. In our time each engineer, scientific worker or manager must have at least an elementary knowledge of probability theory. Experience shows, however, that probability theory is rather difficult for beginners. Those whose formal training does not extend beyond traditional scientific methods find it difficult to adapt to its specific features. And it is these first steps in understanding and application of probability laws that turn out to be the most difficult. The earlier this kind of psychological barrier is overcome the better.

Without claiming a systematic presentation of probability theory, in this small book we shall try to facilitate the reader's first steps. Despite the released (sometimes even humorous) form of presentation, the book at times will demand serious concentration from the reader.

First of all a few words about "random events". Imagine an *experiment* (or a "trial") with results that cannot be predicted. For example, we toss a coin; it is impossible to say in advance whether it will show heads or tails. Another example: we take a card from a deck at random. It is impossible to say what will be its suit. One more example: we come up to a bus stop without knowing the time-table. How long will we have to wait for our bus? We cannot say beforehand. It depends, so to speak, on chance. How many articles will be rejected · by the plant's quality control department? We also cannot predict that.

All these examples belong to random events. In each of them the outcome of the experiment cannot be predicted beforehand. If such experiments with unpredictable outcomes are repeated several times, their successive results will differ. For instance, if a body has been weighed on precision scales, in general we shall obtain different values for its weight. What is the reason for this? The reason is that, although the conditions of the experiments seem quite alike, in fact they differ from experiment to experiment. The outcome of each experiment depends on numerous minor factors which are difficult to grasp and which account for the uncertainty of outcomes.

Thus let us consider an experiment whose outcome is unknown beforehand, that is, random. We shall call a *random event* any event which either happens or fails to happen as the result of an experiment. For example, in the experiment consisting in the tossing of a coin

event A — the appearance of heads — may (or may not) occur. In another experiment, tossing of two coins, event B which can (or cannot) occur is the appearance of two heads simultaneously. Another example. In the Soviet Union there is a very popular game — sport lottery. You buy a card with 49 numbers representing conventionally different kinds of sports and mark any six numbers. On the day of the drawing, which is broadcast on TV, 6 balls are taken out of the lottery drum and announced as the winners. The prize depends on the number of guessed numbers (3, 4, 5 or 6). Suppose you have bought a card and marked 6 numbers out of 49. In this experiment the following events may (or may not) occur:

A — three numbers from the six marked in the card coincide with the published winning numbers (that is, three numbers are guessed),

B — four numbers are guessed,

C — five numbers are guessed, and finally, the happiest (and the most unlikely) event:

D — all the six numbers are guessed.

And so, probability theory makes it possible to determine the degree of likelihood (probability) of various events, to compare them according to their probabilities and, the main thing, to *predict* the outcomes of random phenomena on the basis of probabilistic estimates.

"I don't understand anything!" you may think with irritation. "You have just said that random events are unpredictable. And now — 'we can predict'!"

Wait a little, be patient. Only those random events can be predicted that have a high degree of likelihood or, which comes to the same thing, high probability. And it is probability theory that makes it possible to determine which events belong to the category of highly likely events.

Let us discuss the probability of events. It is quite obvious that not all random events are equally probable and that among them some are more probable and some are less probable. Let us consider the experiment with a die. What do you think, which of the two outcomes of this experiment is more probable:

A — the appearance of six dots

or

B — the appearance of an even number of dots?

If you cannot answer this question at once, that's bad. But the event that the reader of a book of this kind cannot answer such a simple question is highly unlikely (has a low probability!). On the contrary, we can guarantee that the reader will answer at once: "No doubt, event B is more probable!" And he will be quite right, since the elementary understanding of the term "the probability of an event" is inherent in any person with common sense. We are surrounded by random phenomena, random events and since a child, when planning our actions, we used to estimate the probabilities of events and distinguish among them the probable, less probable and impossible events. If the probability of an event is rather low, our common sense tells us not to expect seriously that it will occur. For instance, there are 500 pages in a book, and the formula we need is on one of them. Can we seriously expect that when we open the book at random we will come across this very page? Obviously not. This event is possible but unlikely.

Now let us decide how to evaluate (estimate) the probabilities of random events. First of all we must choose a unit of measurement. In probability theory such a unit is called the *probability of a sure event.* An event is called a sure event if it will certainly occur in the given experiment. For instance, the appearance of

not more than six spots on the face of a die is a sure event. The probability of a sure event is assumed equal to one, and zero probability is assigned to an *impossible event*, that is the event which, in the given experiment, cannot occur at all (e. g. the appearance of a negative number of spots on the face of a die).

Let us denote the probability of a random event A by $P(A)$. Obviously, it will always be in the interval between zero and one:

$$0 \leqslant P(A) \leqslant 1. \tag{1.1}$$

Remember, this is the most important property of probability! And if in solving problems you obtain a probability of greater than one (or, what is even worse, negative), you can be sure that your analysis is erroneous. Once one of my students in his test on probability theory managed to obtain $P(A) = 4$ and wrote the following explanation: "this means that the event is more than probable". (Later he improved his knowledge and did not make such rough mistakes.)

But let us return to our discussion. Thus, the probability of an impossible event is zero, the probability of a sure event is one, and the probability $P(A)$ of a random event A is a certain number lying between zero and one. This number shows which *part* (or share) of the probability of a sure event determines the probability of event A.

Soon you will learn how to determine (in some simple problems) the probabilities of random events. But we shall postpone this and will now discuss some principal questions concerning probability theory and its applications.

First of all let us think over a question: Why do we need to know how to calculate probabilities?

Certainly, it is interesting enough in itself to be able to estimate the degrees of likelihood for various

events and to compare them. But our goal is quite different — using the calculated probabilities to *predict* the outcomes of experiments associated with random events. There exist such experiments whose outcomes are predictable despite randomness. If not precisely, then at least approximately. If not with complete certainty, then with an *almost complete*, that is, "practical certainty". One of the most important problems of probability theory is to reveal the events of a special kind — the *practically sure* and *practically impossible events.*

Event A is called practically sure if its probability is not exactly equal but very close to one:

$$P(A) \approx 1.$$

Similarly, event A is called practically impossible if its probability is close to zero:

$$P(A) \approx 0.$$

Consider an example: the experiment consists in 100 persons tossing coins. Event A corresponds to the appearance of heads simultaneously in all cases. Is this event possible? Theoretically, yes, it is. We can imagine such a "freak of fortune". However, the probability of this event is negligibly small [later we shall calculate it and see that it is $(1/2)^{100}$]. We may conclude that event A can be considered as practically impossible. The opposite event $\bar{A}$ consisting in the fact that A does not occur (that is tails appears at least once) will be practically certain.

In the problems on probability theory practically impossible and practically sure events always appear in pairs. If event A in the given experiment is practically sure, the opposite event $\bar{A}$ is practically impossible, and vice versa.

Suppose we have made some calculations and

established that in the given experiment event A is practically sure. What does this mean? It means that we can *predict* the occurrence of event A in this experiment! True, we can do it not absolutely for sure but "almost for sure", and even this is a great achievement since we are considering random events.

Probability theory makes it possible to predict with a certain degree of confidence, say, the maximum possible error in calculations performed by a computer; the maximum and minimum number of spare parts needed for a lorry fleet per year; the limits for the number of hits of a target and for the number of defective articles in a factory, and so on.

It should be noted that such predictions, as a rule, are possible when we are considering not a single random event but a *great number* of similar random events. It is impossible, for example, to predict the appearance of heads or tails when tossing a coin — probability theory is of no use here. But if we repeat a great number of coin tossings (say, 500 or 1000), we can predict the limits for the number of heads (the examples of similar predictions will be given below, all of them are made not for sure but "almost for sure" and are realized not literally always but in the majority of cases).

Sometimes the question arises: What must be the probability so as to consider an event practically sure? Must it be, say, 0.99? Or, perhaps, 0.995? Or higher still, 0.999?

We cannot answer this question in isolation. All depends on the consequences of occurrence or nonoccurrence of the event under consideration.

For example, we can predict, with a probability of 0.99, that using a certain kind of transport we shall not be late to work by more than 10 min. Can we consider the event practically sure? I think, yes, we can.

But can we consider the event of the "successful landing of a spaceship" to be practically sure if it is guaranteed with the same probability of 0.99? Obviously, not.

Remember that in any prediction made by using the methods of probability theory there are always two specific features.

1. The predictions are made not for sure but "almost for sure", that is, with a high probability.

2. The value of this high probability (in other words, the "degree of confidence") is set by the researcher himself more or less arbitrarily but according to common sense, taking into account the importance of the successful prediction.

If after all these reservations you haven't been completely disappointed in probability theory, then go on reading this book and become acquainted with some simple methods for calculating the probabilities of random events. You probably already have some idea of these methods.

Please answer the question: What is the probability of the appearance of heads when we toss a coin?

Almost for certain (with a high probability!) you will immediately answer: 1/2. And you will be right if the coin is well balanced and the outcome "the coin stands on edge" is rejected as practically impossible. (Those who will think before giving the answer "1/2" are just cavillers. Sometimes it is an indication of deep thinking but most often, of a pernickety nature.)

You can also easily answer another question: What is the probability of the appearance of six spots when rolling a die? Almost for sure your answer will be "1/6" (of course, with the same reservation that the die is fair and that it is practically impossible for it to stop on edge or on a corner).

How did you arrive at this answer? Obviously, you

counted the number of possible outcomes of the experiment (there are six). Due to symmetry the outcomes are equally likely. It is natural to ascribe to each of them a probability of 1/6, and if you did so you were quite right.

And now one more question: What is the probability of the appearance of more than four spots in the same experiment? I imagine your answer will be "1/3". If so, you are right again. Indeed, from six equally possible outcomes two (five and six spots) are said "to imply" this event. Dividing 2 by 6 you get the correct answer, 1/3.

Bravo! You have just used, without knowing it, *the classical model for calculating probability.*

And what in fact is the classical model? Here is the explanation.

First let us introduce several terms (in probability theory, just as in many other fields, terminology plays an important role).

Suppose we are carrying out an experiment which has a number of possible outcomes: A_1, A_2, ..., A_n.

A_1, A_2, ..., A_n are called *mutually exclusive* if they exclude each other, that is, among them there are no two events that can occur simultaneously.

Events A_1, A_2, ..., A_n are called *exhaustive* if they cover all possible outcomes, that is, it is impossible that none of them occurred as a result of the experiment.

Events A_1, A_2, ..., A_n are called *equally likely* if the conditions of the experiment provide equal possibility (probability) for the occurrence of each of them.

If events A_1, A_2, ..., A_n possess all the three properties, that is, they are (a) mutually exclusive; (b) exhaustive and (c) equally likely, they are described by the *classical model.* For brevity we shall call them *chances.*

For example, the experiment "tossing a coin" can be described by the classical model since event A_1 — the appearance of heads and event A_2 — the appearance of tails are mutually exclusive, exhaustive and equally likely, that is, they form a group of chances.

The experiment "casting a die" is also described by the classical model; there are six mutually exclusive, exhaustive and equally likely outcomes which can be denoted according to the number of spots on a face: A_1, A_2, A_3, A_4, A_5, and A_6.

Consider now the experiment consisting in "tossing two coins" and try to enumerate all the possible outcomes. If you are thoughtless and hasty, you will push into suggesting three events:

B_1 — the appearance of two heads;

B_2 — the appearance of two tails;

B_3 — the appearance of one head and one tail.

If so, you will be wrong! These three events are not chances. Event B_3 is doubly more probable than each of the rest. We can verify it if we list the real chances of the experiment:

A_1 — head on the first coin and head on the second;

A_2 — tail on the first coin and tail on the second;

A_3 — head on the first coin and tail on the second;

A_4 — tail on the first coin and head on the second.

Events B_1 and B_2 coincide with A_1 and A_2. Event B_3, however, includes alternatives, A_3 and A_4, and for this reason it is doubly more probable than either of the remaining events.

In the following example we shall for the first time use the traditional model in probability theory — the *urn model*. Strictly speaking, the urn is a vessel containing a certain number of balls of various colours. They are thoroughly mixed and are the same to the touch, which ensures equal probability for any of them to be drawn

out. These conditions will be understood to hold in all the urn problems considered below.

Every problem in probability theory, in which the experiment can be described by the classical model, can be represented by a problem consisting in drawing the balls out of an urn. The urn problems are a kind of unique language into which various problems in probability theory can be translated.

For example, suppose we have an urn with three white and four black balls. The experiment consists in one ball being drawn out at random. Give all possible outcomes of the experiment.

Solving this problem you can again make a mistake and name hastily two events: B_1 — the appearance of a white ball and B_2 — the appearance of a black ball. If you have such an intention you are not born for probability theory. By now it is more likely that you will not give such an answer, since you have understood that in the given experiment there are not two but seven possible outcomes (by the number of balls), which can be denoted, for example, like this: W_1, W_2, W_3, B_1, B_2, B_3 and B_4 (white one, white two, etc., black four). These outcomes are mutually exclusive, exhaustive and equally likely, that is, this experiment can also be described by the classical model.

The question arises: Is it possible to describe any experiment using the classical model? No, of course not. For example, if we are tossing an unbalanced (bent) coin, the events "the appearance of heads" and "the appearance of tails" cannot be described by the classical model since they are not two equally possible outcomes (we could manage to bend the coin in such a way that one of the events would become impossible!). For the experiment to be described by the classical model it must possess a certain symmetry which would provide equal probabilities for all possible

outcomes. Sometimes this symmetry can be achieved due to the physical symmetry of the objects used in the experiment (a coin, a die), and sometimes due to mixing or shuffling the elements, which provides an equal possibility for any of them to be chosen (an urn with balls, a deck of cards, a lottery drum with tickets, etc.). Most frequently such a symmetry is observed in artificial experiments in which special measures are taken to provide the symmetry. Typical examples are games of chance (e.g. dice and some card games). It should be noted that it was the analysis of games of chance that gave the impulse for the development of probability theory.

If an experiment can be described by the classical model, the probability of any event A in this experiment can be calculated as *the ratio of the number of chances favourable for event A to the total number of chances*:

$$P(A) = \frac{m_A}{n} \tag{1.2}$$

where n is the total number of chances and m_A is the number of chances favourable for event A (that is, implying its appearance).

Formula (1.2), the so-called *classical formula*, has been used for the calculation of the probabilities of events since the very onset of the science of random events. For a long time it was assumed to be the *definition of probability*. The experiments which didn't possess the symmetry of possible outcomes were artificially "drawn" into the classical model. In our time the definition of probability and the manner of presentation of probability theory have changed. Formula (1.2) is not general but it will enable us to calculate the probabilities of events in some simple cases. In the following chapters we will learn how to

calculate the probabilities of events when the experiment cannot be described by the classical model.

Consider now several examples in which we can calculate the probabilities of random events using formula (1.2). Some of them are very simple but the others are not.

EXAMPLE 1.1. The experiment consists in throwing two coins. Find the probability of the appearance of heads at least once.

Solution. Denote by A the appearance of heads at least once. There are four chances in this experiment (as was mentioned on p. 15). Three of them are favourable for event A (A_1, A_3, and A_4). Hence, $m_A = 3$, $n = 4$, and formula (1.2) yields

$$P(A) = 3/4.$$

EXAMPLE 1.2. The urn contains three white and four black balls. One of the balls is drawn out of the urn. Find the probability that the ball is white (event A).

Solution. $n = 7$, $m_A = 3$, $P(A) = 3/7$.

EXAMPLE 1.3. The same urn with three white and four black balls, but the conditions of the experiment are somewhat changed. We take out a ball and put it into a drawer without looking at it. After that we take out a second ball. Find the probability that this ball will be white (event A).

Solution. If you think a little you will see that the fact that we have taken the ball of unknown colour out of the urn in no way affects the probability of the appearance of a white ball in the second trial. It will remain the same as in Example 1.2: 3/7.

On the first thought the result may seem incorrect, since before we took out the second ball, instead of seven there were six balls in the urn. Does this mean that the total number of chances became six?

No, it doesn't. Unless we knew what colour the first

ball was (for this purpose we hid it in the drawer!) the total number of chances n remains seven. To convince ourselves of this let us alter the conditions of the experiment still more radically. This time we draw out all the balls but one and put them into the drawer without looking at them. What is the probability that the last ball is white? Clearly, $P(A) = 3/7$ since it is absolutely the same whether the ball is *drawn out* or *left alone* in the urn.

If you are still not convinced that the probability of the appearance of a white ball is 3/7 regardless of the number of balls of unknown colours placed beforehand in the drawer, imagine the following experiment. We have 3 white and 4 black balls in the urn. It is dark in the room. We draw the balls out of the urn and spread them around the room: two on the window-sill, two on the cupboard, one on the sofa, and the remaining two throw on the floor. After that we start walking about the room and step on a ball. What is the probability that the ball is white?

If you still cannot understand why it is 3/7, nothing will help you; all our arguments are exhausted.

EXAMPLE 1.4. The same urn with 3 white and 4 black balls. If two balls are drawn out simultaneously, what is the probability that both balls will be white (event A)?

Solution. This problem is a little more difficult than the two previous problems; in this case it is more difficult to calculate the total number of chances n and the number of favourable chances m_A. Here we have to find the number of possible ways of selecting two balls from the urn and the number of ways of selecting two balls out of the white balls.

The number of combinations for selecting and arranging some elements from the given set of elements can be calculated by the methods of combinatorial

analysis—a discipline which is part of elementary algebra. Here we shall require only one formula of combinatorial analysis—the formula for the number of combinations $C(k, s)$. We remind you that the number of combinations of k elements taken s at a time is the number of ways of selecting s different elements out of k (combinations differ only in the composition of the elements and not in their order). The number of combinations of k elements taken s at a time is calculated by the formula

$$C(k,s) = \frac{k(k-1)\ldots(k-s+1)}{1\cdot 2\cdot \ldots s} \tag{1.3}$$

and has the following property:

$$C(k,s) = C(k,k-s). \tag{1.4}$$

Let us calculate, using formula (1.3), the number n of all possible chances in our example (the number of ways in which we can select two balls out of seven):

$$n = C(7,2) = \frac{7\cdot 6}{1\cdot 2} = 21.$$

Now we shall calculate the number of favourable chances m_A. This is the number of ways in which we can select two balls out of three white balls in the urn. Using formulas (1.4) and (1.3), we find

$$m_A = C(3,2) = C(3,1) = 3,$$

whence by formula (1.2) we obtain

$$P(A) = \frac{C(3,1)}{C(7,2)} = \frac{3}{21} = \frac{1}{7}.$$

EXAMPLE 1.5. Still the same urn (be patient, please, we shall soon be finishing with it!) with 3 white and 4 black balls. Three balls are taken out simultaneously.

What is the probability that two of them are black and one white (event A)?

Solution. Calculate the total number of chances in the given experiment:

$$n = C(7,3) = \frac{7 \cdot 6 \cdot 5}{1 \cdot 2 \cdot 3} = 35.$$

Let us calculate the number of favourable chances m_A. In how many ways can we select two out of four black balls? Clearly, $C(4,2) = \frac{4 \cdot 3}{1 \cdot 2} = 6$ ways. But this is not all; to each combination of two black balls we can add one white ball in different ways whose number is $C(3,1) = 3$. Every combination of black balls can be combined with one of white balls, therefore, the total number of favourable chances is

$$m_A = C(4,2) \cdot C(3,1) = 6 \cdot 3 = 18,$$

whence by formula (1.2) we have

$$P(A) = 18/35.$$

Now we are skilled enough to solve the following general problem.

PROBLEM. There are a white and b black balls in an urn; k balls are drawn out at random. Find the probability that among them there are l white and, hence, $k-l$ black balls ($l \leqslant a$, $k - l \leqslant b$).

Solution. The total number of chances

$$n = C(a + b,\ k).$$

Calculate the number of favourable chances m_A. The number of ways in which we can select l white balls out of a balls is $C(a, l)$; the number of ways to add $k - l$ black balls is $C(b,\ k - l)$. Each combination of white balls can be combined with each combination of black

balls; therefore

$$P(A) = \frac{C(a,\ l)\,C(b,\ k-l)}{C(a+b,\ k)}. \tag{1.5}$$

Formula (1.5) is widely applied in various fields, for example, in the problems concerning the selective control of manufactured items. Here the batch of manufactured items is considered as an urn with a certain number of nondefective items (white balls) and a certain number of defective items (black balls). The k items selected for examination play the role of balls which are drawn out of the urn.

After solving the general problem let us consider one more example which you may find interesting.

EXAMPLE 1.6. Somebody has bought a card in the sport lottery and marked six numbers out of the 49. What is the probability that he guessed three numbers which will be among the six winning numbers?

Solution. Consider event A which consists in 3 numbers out of 6 having been guessed (this means that 3 numbers have not been guessed).

Just a minute! This is exactly our previous problem! Indeed, 49 numbers with 6 winning ones can be represented by the urn with 6 white and 43 black balls. We have to calculate the probability that taking 6 balls out of the urn at random we shall have 3 white and 3 black balls. And we know how to do it! In formula (1.5) make $a = 6$, $b = 43$, $k = 6$, and $l = 3$ and find

$$P(A) = \frac{C(6,3)\cdot C(43,3)}{C(49,6)} = \frac{6\cdot 5\cdot 4\cdot 43\cdot 42\cdot 41\cdot 4\cdot 5}{49\cdot 48\cdot 47\cdot 46\cdot 45\cdot 44}.$$

If you have the energy and take the trouble to calculate the result you will get $P(A) \approx 0.0176$.

Well, the probability that you guessed three numbers out of six is rather small—about 1.8%. Obviously, the probability of guessing four, five and (oh, wonder!) all six numbers is still lower. We can say, considerably lower. If you are fond of calculations you may try and calculate the probabilities of these events. So, you have already got somewhere... .

Probability and Frequency

In Chapter 1 you have been acquainted with the subject matter of probability theory, with its principal concepts (an event and the probability of an event) and learned how to calculate the probabilities of events using the so-called classical formula

$$P(A) = \frac{m_A}{n}, \tag{2.1}$$

where n is the total number of chances, and m_A is the number of chances favourable for event A.

This does not mean, however, that you are able to apply probability theory in practice. Formula (2.1) is unfortunately not as general as could be desired. We can use it only in experiments that possess a symmetry of possible outcomes (that is, which can be described by the classical model). These experiments are mainly games of chance where symmetry is ensured by special measures. And since in our days gambling is not a widespread profession (in past times the situation was different), the practical importance of formula (2.1) is very limited. Most of the experiments with random

results which we deal with in practice cannot be described by the classical model. What can we say about the probabilities of events? Do they exist for such experiments? And if they do then how do we find them?

That is the problem we shall consider now. Here we have to introduce a new concept of probability theory — the *frequency of an event.*

We shall approach it, so to say, from afar. Imagine an experiment which cannot be described by the classical model, for instance, rolling an irregular asymmetric die (the asymmetry can be achieved, for instance, by the displacement of the die's center of mass with the help of a lead load which increases the probability of the appearance of a certain face). This kind of trick was used by professional gamblers in their time for gain. (Incidentally, it was a rather dangerous business, for if a gambler was caught redhanded he was usually severely beaten and sometimes to death.) Consider in this experiment an event A — the appearance of six spots. Since the experiment cannot be described by the classical model, formula (2.1) is of no use here and we cannot assume that $P(A) = 1/6$. Then what is the probability of event A? Is it higher or lower than 1/6? And how can it be determined, at least approximately? Any person with common sense can easily answer this question. He would say: "You should *try* and cast the die many times and see how *often* (as a proportion of the trials) event A takes place. The fraction (or, in other terms, percent) of occurrences of event A can be taken as its probability.

Well, this person "with common sense" is absolutely right. Without knowing it he has used the concept of frequency of an event. Now we will give the exact definition of this term.

The frequency of an event in a series of N repetitions

is the ratio of the number of repetitions, in which this event took place, to the total number of repetitions.

The frequency of an event is sometimes called its *statistical probability*. It is the statistics of a great number of random events that serves as the basis for determining the probability of events in experiments possessing no symmetry with respect to possible outcomes.

The frequency of event A will be denoted by $P^*(A)$ (the asterisk is used to distinguish the frequency from the probability $P(A)$ related to it). By definition,

$$P^*(A) = \frac{M_A}{N}, \tag{2.2}$$

where N is the total number of repetitions of the experiment and M_A is the number of repetitions in which event A occurs (in other words, the number of occurrences of event A).

Despite their similarity in form, formulas (2.1) and (2.2) are essentially different. Formula (2.1) is used in *theoretical calculation* of the probability of an event according to the given conditions of the experiment. Formula (2.2) is used for the *experimental determination* of the frequency of an event; to apply it we need experimental statistical data.

Let us think about the nature of frequency. It is quite clear that there is a certain relation between the frequency of an event and its probability. Indeed, the most probable events occur more frequently than events with a low probability. Nevertheless, the concepts of frequency and probability are not identical. The greater is the number of repetitions of the experiment the more marked is the correlation between the frequency and the probability of the event. If the number of repetitions is small, the frequency of the

event is to a considerable extent a random quantity which can essentially differ from the probability. For example, we have tossed a coin 10 times and heads appeared 3 times. This means that the frequency of this event is 0.3, and this value considerably differs from the probability of 0.5. However, with the increasing number of repetitions the frequency of the event gradually loses its random nature. Random variations in the conditions of each trial cancel each other out when the number of trials is great; the frequency tends to stabilize and gradually approaches, with small deviations, a certain constant. It is natural to suppose that this constant is exactly the *probability* of the event.

Of course, this assertion can be verified only for those events whose probabilities can be calculated by formula (2.1), that is, for the experiments which can be described by the classical model. It turns out that for those cases the assertion is true.

If you are interested you can verify it for yourself in some simple experiment. For example, see for yourself that as the number of tosses of a coin increases the frequency of the appearance of a head approaches the probability of this event of 0.5. Throw the coin 10, 20, ... times (till you lose patience) and calculate the frequency of the appearance of a head depending on the number of tosses. To save yourself time and effort you can resort to cunning and throw not one coin but ten coins at a time (of course you must mix them thoroughly beforehand). Then plot a graph according to the obtained figures. You will obtain a curve of the kind shown in Fig. 1, which was the result of tossing thoroughly mixed 10 one-copeck coins 30 times. Perhaps you are more patient and are able to make a much greater number of tosses. It might be of interest to know that even eminent scientists studying the nature of random events didn't ignore such

experiments. For example, the famous statistician Karl Pearson tossed a coin 24,000 times and obtained 12,012 heads, which corresponds very closely to the frequency of 0.5. The experiment of casting a die has also been repeated a great number of times and resulted in the frequencies of appearance of different faces close to 1/6.

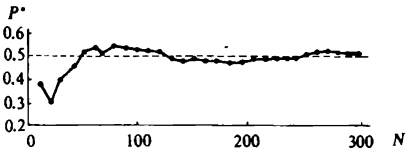


Fig. 1

Thus, the fact that frequencies tend to respective probabilities can be considered experimentally verified.

Stabilization of frequency with a great number of repetitions of an experiment is one of the most typical properties observed in mass random phenomena. If we repeat (reproduce) one and the same experiment many times (provided the outcomes of individual trials are *mutually independent*) the frequency of the event becomes less and less random; it levels out and approaches a constant. For the experiments described by the classical model we can directly make sure that this constant is exactly the *probability of the event.* And what if the experiment cannot be reduced to the classical model but the frequency exhibits stability and tends to a constant? Well, we shall assume that the rule remains in force and identify this constant with the probability of the event. Thus, we have introduced the concept of probability not only for the events and experiments that can be reduced to the classical model but also for those which cannot be reduced to it but in which *frequency stability* is observed.

And what can we say about the stabilization of fre-

quencies? Do all random phenomena possess this property? Not all of them, but many do. We shall explain this rather complicated idea by the following reasoning. Try to understand it properly; this will rid you of possible mistakes when you use probability methods.

When speaking about frequency stabilization we assumed that we could repeat one and the same experiment (tossing a coin, casting a die, etc.) an unlimited number of times. Indeed, nothing prevents us from making such an experiment however large a number of times (if we have the time). But there are cases when we do not carry out an experiment ourselves but just observe the results of "experiments" carried out by nature. In such cases we cannot guarantee the frequency stabilization beforehand, we must be convinced of it each time.

Suppose, for example, the "experiment" is the birth of a child, and we would like to know the probability of a boy being born ("experiments" of this kind are made by nature in vast numbers annually). Is there frequency stabilization in these random events? Yes, there is. It is established statistically that the frequency of the birth of a boy is very stable and almost independent of the geographical location of the country, the nationality of parents, their age, etc. This frequency is somewhat higher than 0.5 (it is approximately 0.51). In some special conditions (for example, during war or after it) this frequency, for reasons unknown so far, can deviate from the average stable value observed for many years.

The property of frequency stabilization (at least for fairly short periods) is inherent in such random phenomena as failure of technical devices, rejects in production, errors of mechanisms, morbidity and mortality of population, and meteorological and

biological phenomena. This stabilization enables us to apply successfully probability methods for studying such random phenomena, to predict and control them.

But there are also random phenomena for which frequency stabilization is dubious, if it exists at all. These are such phenomena for which it is meaningless to speak of a great number of repetitions of an experiment and for which there is no (or cannot be obtained in principle) sufficient statistical data. For such phenomena there are also certain events which seem to us more or less probable, but it is impossible to ascribe to them definite probabilities. For example, it is hardly possible (and scarcely reasonable!) to calculate the probability that in three years women will wear long skirts (or that for men long moustache will be in fashion). There is no appropriate statistical data file for that; if we consider the succeeding years as "experiments", they cannot be taken as similar in any sense.

Here is another example where it is even more meaningless to speak of the probability of an event. Suppose we decided to calculate the probability that on Mars there exists organic life. Apparently, the question of whether or not it exists will be solved in the near future. Many scientists believe that organic life on Mars is quite plausible. But the "degree of plausibility" is not yet the probability! Estimating the probability "by a rule of thumb" we inevitably find ourselves in the world of vague conjecture. If we want to operate with actual probabilities, our calculations must be based on sufficiently extensive statistics. And can we speak about vast statistical data in this case? Of course not. There is only one Mars!

So, let us clarify our position: we shall speak of the probabilities of events in experiments which cannot be described by the classical model only if they belong to

the type of mass random events possessing the property of frequency stabilization. The question whether an experiment possesses this property or not is usually answered from the standpoint of common sense. Can the experiment be repeated many times without essentially varying its conditions? Is there a chance that we accumulate appropriate statistical data? These questions must be answered by a researcher who is going to apply probability methods in a particular field.

Let us consider another question directly connected with what was said above. Speaking about the probability of an event in some experiment, we must first of all specify thoroughly the basic conditions of the experiment, which are assumed to be fixed and invariable throughout the repetitions of the experiment. One of the most frequent mistakes in the practical application of probability theory (especially for beginners) is to speak of the probability of an event without specifying the conditions of the experiment under consideration, and without the "statistical file" of random events in which this probability could be manifested in the form of frequency.

For example, it is quite meaningless to speak of the probability of such an event as the delay of a train. At once a number of questions arise: What train is meant? Where does it come from and what is its destination? Is it a freight train or a passenger train? Which railway does it belong to? and so on. Only after clarifying all these details can we consider the probability of this event as a definite value. So we must warn you about the "reefs" that endanger those who are interested not only in the problems for fun about "coins", "dice", and "playing cards" but who want to apply the methods of probability theory in practice to achieve real goals.

Now suppose that all these conditions are satisfied, that is, we can make a sufficient number of similar trials and the frequency of an event is stable. Then we can use the frequency of the event in the given series of trials as an approximation of the probability of this event. We have already agreed that as the number of trials increases, the frequency of the event approaches its probability. It seems that there is no problem but in fact it isn't as simple as that. The relationship between frequency and probability is rather intricate.

Think a little about the word "approaches" that we have used above. What does it mean?

"What a strange question," you may think. "'Approaches' means that it becomes closer and closer. So what do we have to think about?"

Still, there is something to think about. You see, we are dealing with random phenomena, and with these phenomena everything is peculiar, "eccentric".

When we say that the sum of geometrical progression

$$1 + \frac{1}{2} + \frac{1}{2^2} + \ldots + \frac{1}{2^n}$$

with an increasing n tends to (infinitely approaches) 2, this means that the more terms we take the closer is the sum to its limit, and this is absolutely true. In the field of random phenomena it is impossible to make such categorical assertions. Indeed, in general the frequency of an event tends to its probability but in its own manner: not for certain, but *almost for certain*, with a very high probability. But it can happen that even with a very large number of trials the frequency of the event will differ considerably from the probability. The probability of this, however, is very low, and it is the lower the greater the number of trials conducted.

Suppose we have tossed a coin $N = 100$ times. Can

it so happen that the frequency of appearance of heads will considerably differ from the probability $P(A) = 0.5$, for example, will be zero? To this end it is required that heads do not appear at all. Such an event is theoretically possible (it does not contradict the laws of nature), but its probability is extremely low. Indeed, let us calculate this probability (fortunately, we can now solve such an easy problem). Calculate the number of possible outcomes n. Each toss of a coin can result in two outcomes: a head or a tail. Each of these two outcomes of one toss can be combined with any of two outcomes of other tosses. Hence the total number of possible outcomes is 2^{100}. Only one outcome is favourable to our event (not a single head); it means that its probability is $1/2^{100}$. This is a very small value; in the order of 10^{-30}, that is, the number representing this value contains 30 noughts after the decimal point. The event with such a probability can surely be considered practically impossible. In reality even much smaller deviations of frequency from probability are practically impossible.

What deviations of frequency from probability are practically possible when the number of trials is great? Now we shall give the formula which makes it possible to answer this question. Unfortunately, we cannot prove this formula (to a certain extent its justification will be made later); so at present you must take it for granted. They say that the famous mathematician d'Alembert, giving lessons in mathematics to a very noble and very stupid student, couldn't explain to him the proof of a theorem and desperately exclaimed: "Upon my word, sir, this theorem is correct!" His pupil replied: "Sir, why didn't you say so earlier? You are a noble man and I am too. Your word is quite enough for me!"

Suppose N trials are made, in each of them an event

A appears with probability p. Then, with the probability of 0.95, the value of frequency $P^*(A)$ of event A lies in the interval

$$p \pm 2\sqrt{\frac{p(1-p)}{N}}. \tag{2.3}$$

The interval determined by formula (2.3) will be called the confidence interval for the frequency of the event, which corresponds to the confidence level of 0.95. This means that our prediction that the value of frequency will be in the limits of this interval is true in almost all cases, to be exact, in 95% of cases. Of course, in 5% of cases we will be mistaken but... he that never climbed never fell. In other words, if you are afraid of errors don't try to predict in the field of random events — the predictions are fulfilled not for sure but almost for sure.

But if the value 0.05 for the probability of error seems too high to you, then to be on the safe side we can take a wider confidence interval:

$$p \pm 3\sqrt{\frac{p(1-p)}{N}}. \tag{2.4}$$

This interval corresponds to a very high confidence level of 0.997.

And if we require the complete reliability of our prediction, that is, confidence coefficient equal to one, what then? In this case we can only assert that the frequency of the event will be in the interval between 0 and 1 — quite a trivial assertion which we could make without calculations.

As was noted in Chapter 1, the probability of the event which we consider as practically sure (that is, the confidence level) is to a certain extent an arbitrary

value. Let us agree that in further estimates of the accuracy in determining probability by frequency we shall be content with a moderate value for the confidence level of 0.95 and use formula (2.3). If we sometimes make a mistake nothing terrible will happen.

Let us estimate, using formula (2.3), the practically possible interval of values (confidence interval) for the frequencies of the appearance of heads on 100 tosses of a coin. In our experiment $p = 0.5$, $1 - p = 0.5$, and formula (2.3) yields

$$0.5 \pm 2\sqrt{\frac{0.25}{100}} = 0.5 \pm 0.1.$$

Thus, with the probability (confidence level) of 0.95 we can predict that on 100 tosses of a coin the frequency of appearance of a head will not differ from the respective probability by more than 0.1. Frankly speaking, the error is not small. What can we do to decrease it? Obviously, we must increase the number N of trials. With increasing N the confidence interval decreases (unfortunately, not as quickly as we desire, in inverse proportion to $\sqrt{N}$). For example, for $N = 10{,}000$ formula (2.3) gives 0.5 ± 0.01.

Consequently, the relationship between the frequency and the probability of an event can be formulated as follows:

If the number of independent trials is sufficiently large, then with a practical confidence the frequency will be as close to the probability as desired.

This statement is called the Bernoulli theorem, or the simplest form of the law of large numbers. Here we give it without proof; however, you can hardly entertain serious doubts about its validity.

Thus, we have elucidated the meaning of the

expression "the frequency of an event tends to its probability". Now we must take the next step—approximately find the probability of the event by its frequency and estimate the error of this approximation. This can be done still using the same formula (2.3) or, if you wish, (2.4).

Suppose we have made a large number N of trials and found the frequency $P^*(A)$ of event A, and now we want to find approximately its probability. Denote for brevity the frequency $P^*(A) = p^*$ and the probability $P(A) = p$ and assume that the sought probability is approximately equal to the frequency:

$$p \approx p^*. \tag{2.5}$$

Now let us estimate the maximum practically possible error in (2.5). For this we shall use formula (2.3), which will help us calculate (with a confidence level of 0.95) the maximum possible difference between the frequency and the probability.

"But how?" you may ask, "formula (2.3) includes the unknown probability p which we want to determine!"

Quite a reasonable question. You are quite right! But the point is that formula (2.3) gives only an *approximate estimate* of the confidence interval. In order to estimate *roughly* the error in the probability, in (2.3) we substitute the known frequency p^* for the unknown probability p which is approximately equal to it.

Let us do it this way! For example, let us solve the following problem. Suppose in a series of $N = 400$ trials the frequency $p^* = 0.25$ has been obtained. Find, with the confidence level of 0.95, the maximum practically possible error in the probability if we make it equal to the frequency of 0.25.

By formula (2.3) approximately replacing p by $p^* = 0.25$, we obtain

$$0.25 \pm 2\sqrt{\frac{0.25 \cdot 0.75}{400}}$$

or approximately 0.25 ± 0.043.

So, the maximum practically possible error is 0.043.

And what must we do if such an accuracy is not high enough for us? If we need to know the probability with a greater accuracy, for example, with the error, say, not exceeding 0.01? Clearly, we must increase the number of trials. But to what value? To establish this we again apply our favourite formula (2.3). Making the probability p approximately equal to frequency $p^* = 0.25$ in the series of trials already made, we obtain by formula (2.3) the maximum practically possible error

$$2\sqrt{\frac{0.25 \cdot 0.75}{N}}.$$

Equate it to the given value of 0.01:

$$2\sqrt{\frac{0.25 \cdot 0.75}{N}} = 0.01.$$

By solving this equation for N we obtain $N = 7500$.

Thus, in order to calculate, by frequency, the probability of the order of 0.25 with the error not exceeding 0.01 (with the confidence level of 0.95), we must make (Oh, boy!) 7500 experiments.

Formula (2.3) [or (2.4) similar to it] can help us answer one more question: Can the deviation of the frequency from the probability be explained by *chance factors* or does this deviation indicate that the *probability is not such as we think?*

For example, we toss a coin $N = 800$ times and obtain the frequency of the appearance of heads equal to 0.52. We suspect that the coin is unbalanced and heads appear more frequently than tails. Is our suspicion justified?

To answer this question we shall start from the assumption that everything is in order: the coin is balanced, the probability of the appearance of heads is normal, that is 0.5, and find the confidence interval (for the confidence level of 0.95) for the frequency of the event "appearance of heads". If the value of 0.52 obtained as the result of the experiment lies within this interval, all is well. If not, the "regularity" of the coin is under suspicion. Formula (2.3) approximately gives the interval 0.5 ± 0.035 for the frequency of appearance of heads. The value of the frequency obtained in our experiment fits into this interval. This means that the regularity of the coin is above suspicion.

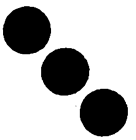
Similar methods are used to judge whether the deviations from the mean value observed in random phenomena are accidental or "significant" (for example, whether the underweight of some package goods is random or it indicates the scales fault; whether the increase in the number of cured patients is due to a medicine they used or it is random).

Thus, we have learned how to find the approximate probability of an event by statistical data in experiments which cannot be described by the classical model, and we can even estimate the possible error.

"What a science probability theory is!" you may think. "If we cannot calculate the probability of an event by formula (2.1), we have to make an experiment and repeat it till we are exhausted, then we must calculate the frequency of the event and make it equal to the probability, after which, perhaps, we shall be in a position to estimate the possible error. What a bore!"

If this is what you think you are absolutely wrong! The statistical method of determining probability is not the only one and is far from being the basic method. In probability theory indirect methods have considerably wider application than direct methods. Indirect methods enable us to represent the probability of events we are interested in by the probabilities of other events related to them, that is, the probabilities of complex events through the probabilities of simple events, which in turn can be expressed in terms of the probabilities of still simpler events, and so on. We can continue this procedure until we come to simplest events the probabilities of which can either be calculated by formula (2.1) or found experimentally by their frequencies. In the latter case we obviously must conduct experiments or gather statistical data. We must do our best to make the chain of the events as long as possible and the required experiment as simple as possible. Of all the materials required to get information the time the researcher spends and the paper he uses are the cheapest. So we must try to obtain as much information as possible using calculations and not experiments.

In the next chapter we shall consider methods of calculating the probabilities of complex events through the probabilities of simple events.



The Basic Rules of Probability Theory

In the preceding chapter we emphasized that the major role in probability theory is played not by *direct* but by *indirect* methods which make it possible to calculate the probabilities of some events by the probabilities of other, simpler events. In this chapter we shall consider these methods. All of them are based on two main principles, or *rules*, of probability theory: (1) the probability summation rule and (2) the probability multiplication rule.

1. Probability summation rule

The probability that one of two mutually exclusive events (it does not matter which of them) occurs is equal to the sum of the probabilities of these events.

Let us express this rule by a formula. Let A and B be two mutually exclusive events. Then

$$P(A \text{ or } B) = P(A) + P(B). \tag{3.1}$$

You may ask: "Where did this rule come from? Is it a theorem or an axiom?" It is both. It can be strictly

proved for the events described by the classical model. Indeed, if A and B are mutually exclusive events, then the number of outcomes implying the complex event "A or B" is equal to $m_A + m_B$, hence

$$P(A \quad \text{or} \quad B) = \frac{m_A + m_B}{n} = \frac{m_A}{n} + \frac{m_B}{n} = P(A) + P(B).$$

For this reason the summation rule is often called the "summation theorem". But you must not forget that it is a theorem only for the events described by the classical model; in other cases it is accepted without proof as a *principle* or *axiom*. It should be noted that the summation rule is valid for frequencies (statistical probabilities); you can prove this yourself.

The summation rule can be easily generalized for any number of events: *The probability that one of several mutually exclusive events (it does not matter which of them) occurs is equal to the sum of the probabilities of these events.*

This rule is expressed by the formula as follows. If events A_1, A_2, ..., A_n are mutually exclusive, then

$$P(A_1 \text{ or } A_2 \text{ or } \ldots \text{ or } A_n) = P(A_1) + P(A_2) + \ldots + P(A_n). \quad (3.2)$$

The probability summation rule has some important corollaries. First, *if events A_1, A_2, ..., A_n are mutually exclusive and exhaustive, the sum of their probabilities is equal to one.* (Try to prove this corollary.) Second, (the corollary of the first corollary), if A is an event and $\bar{A}$ is the opposite event (consisting in nonoccurrence of the event A), then

$$P(A) + P(\bar{A}) = 1, \quad (3.3)$$

that is, *the sum of the probabilities of opposite events is equal to one.*

Formula (3.3) serves as the basis for the "method of transition to the opposite event", which is widely used in probability theory. It often happens that the probability of a certain event A is difficult to calculate, but it is easy to find the probability of the opposite event $\bar{A}$. In this case we can calculate the probability $P(\bar{A})$ and subtract it from unity:

$$P(A) = 1 - P(\bar{A}). \tag{3.4}$$

2. Probability multiplication rule

The probability of the combination of two events (that is, of the simultaneous occurrence of both of them) is equal to the probability of one of them multiplied by the probability of the other provided that the first event has occurred.

This rule can be described by the formula

$$P(A \text{ and } B) = P(A) \cdot P(B/A), \tag{3.5}$$

where $P(B/A)$ is the so-called *conditional probability* of an event B calculated for the condition (on the assumption) that event A has occurred.

The probability multiplication rule is also a *theorem* which can be strictly proved within the framework of the classical model; in other cases it must be accepted without proof as a *principle* or an *axiom*. For frequencies (statistical probabilities) it is also observed.

In should be noted that when the probability multiplication rule is used, it is all the same which of the events is considered to be "the first" and which "the second". This rule can also be written in the form

$$P(A \text{ and } B) = P(B) \cdot P(A/B). \tag{3.6}$$

EXAMPLE 3.1. Let the urn contain 3 white balls and 4 black balls. Two balls are extracted one after the other. Find the probability that both these balls will be white.

Solution. Consider two events: A — the first ball is white, and B — the second ball is also white. We must find the probability of the combination of these events (both the first and the second balls being white). By the probability multiplication rule we have

$$P(A \text{ and } B) = P(A) \cdot P(B/A),$$

where $P(A)$ is the probability that the first ball will be white and $P(B/A)$ is the conditional probability that the second ball will also be white (calculated on the condition that the first ball is white).

Clearly, $P(A) = 3/7$. Let us calculate $P(B/A)$. If the first ball is white, the second ball is selected from the remaining 6 balls of which two balls are white; hence, $P(B/A) = 2/6 = 1/3$. This yields

$$P(A \text{ and } B) = (3/7) \cdot (1/3) = 1/7.$$

Incidentally, we obtained this result earlier (see Example 1.4) using a different method — by the direct calculation of possible outcomes.

Thus, the probability that two balls extracted one after the other out of the urn are white is 1/7. Now answer the question: Will the solution be different if we extract the balls not one after the other but *simultaneously?* At first glance it may seem that it will. But after some thought you will see that the solution will not change.

Indeed, suppose we draw two balls out of the urn simultaneously but with both hands. Let us agree to call "the first ball" the ball in the right hand and "the second ball" that in the left hand. Will our reasoning differ from that we used in the solution of Example 3.1? Not at all! The probability of the two balls being white remains the same: 1/7. "But if we extract the balls with one hand", somebody may persist. "Well, then we will call 'the first ball' the one which is nearer

to the thumb and 'the second ball' the one that is closer to the littre finger." "But if...," our distrustful questioner cannot stop. We reply: "If you still have to ask... shame on you."

The probability multiplication rule becomes especially simple for a particular type of events which are called *independent*. Two events A and B are called independent if the occurrence of one of them does not affect the probability of the other, that is, the conditional probability of event A on the assumption that event B has occurred is the same as without this assumption:

$$P(A/B) = P(A). \tag{3.7}$$

In the opposite case events A and B are called *dependent**.

In Example 3.1 events A and B are dependent since the probability that the second ball drawn out of the urn will be white depends on whether the first ball was white or black. Now let us alter the conditions of the experiment. Suppose that the first ball drawn out of the urn is returned to it and mixed with the rest of the balls, after which a ball is again drawn out. In such an experiment event A (the first ball being white) and event B (the second ball being white) are independent:

$$P(B/A) = P(B) = 3/7.$$

The concepts of the dependence and independence of events are very important in probability theory.

---

* It is easy to prove that the dependence or independence of events is always mutual, that is, if $P(A/B) = P(A)$, then $P(B/A) = P(B)$. The reader can do it independently using two forms of the probability multiplication rule, (3.5) and (3.6).

Incomplete understanding of these concepts often leads to mistakes. This often happens with beginners who have a tendency to forget about the dependence of events when it exists or, on the contrary, find a dependence in events which are in fact independent.

For example, we toss a coin ten times and obtain heads every time. Let us ask somebody unskilled in probability theory a question: Which is the more probable to be obtained in the next, eleventh trial: heads or tails? Almost for sure he will answer: "Of course, tails! Heads have already appeared ten times! It should be somehow compensated for and tails should start to appear at last!" If he says so he will be absolutely wrong, since the probability that heads will appear in every next trial (of course, if we toss the coin in a proper way, for instance put it on the thumb nail and snap it upwards) is completely independent of the number of heads that have appeared beforehand. If the coin is *fair*, then on the eleventh toss, just as on the first one, the probability that a head appears is 1/2. It's another matter that the appearance of heads ten times running may give rise to doubts about the regularity of the coin. But then we should rather suspect that the probability of the appearance of a head on each toss (including the eleventh!) would be greater and not less than 1/2... .

"It looks as if something is wrong here," you may say. "If heads have appeared ten times running, it can't be that on the eleventh toss the appearance of tails would not be more probable!"

Well, let's argue about this. Suppose a year ago you tossed a coin ten times and obtained ten heads. Today you recalled this curious result and decided to toss the coin once more. Do you still think that the appearance of tails is more probable than that of a head? Perhaps you are uncertain now... .

To clinch the argument suppose that somebody else, say the statistician K.' Pearson, many years ago tossed a coin 24,000 times and on some ten tosses obtained ten heads running. Today you recalled this fact and decided to continue his experiment — to toss a coin once more. Which is more probable: heads or tails? I think you've already surrendered and admitted that these events are equally probable.

Let us extend the probability mutiplication rule for the case of several events $A_1, A_2, \ldots, A_n$. In the general case, when the events are not independent, the probability of the combination of the events is calculated as follows: the probability of one event is multiplied by the conditional probability of the second event provided that the first event has occurred, then, by the conditional probability of the third event provided that the two preceding events have occurred, and so on. We shall not represent this rule by a formula. It is more convenient to memorize it in the verbal form.

EXAMPLE 3.2. Let five balls in the urn be numbered. They are drawn from the urn one after the other. Find the probability that the numbers of the balls will appear in the increasing order.

*Solution**. According to the probability multiplication rule,

$$P(1, 2, 3, 4, 5) = (1/5)\cdot(1/4)\cdot(1/3)\cdot(1/2)\cdot 1 = 1/120.$$

EXAMPLE 3.3. In a boy's schoolbag let there be 8 cut-up alphabet cards with the letters: two a's, three c's

---

* Here and below, to simplify presentation, we shall not introduce letter notation for events and give instead in the parentheses following the probability symbol P a brief and comprehensible form of presentation of the event under study. The less notations the better!

and three t's. We draw out three cards one after the other and put them on the table in the order that they appeared. Find the probability that the word "cat" will appear.

Solution. According to the probability multiplication rule,

$$P(\text{"cat"}) = (3/8)\cdot(2/7)\cdot(3/6) = 3/56.$$

And now let us try to solve the same problem in a different form. All the conditions remain the same but the cards with letters are drawn simultaneously. Find the probability that we obtain the word "cat" from the drawn letters.

"Tell me another," you may think. "I remember that it is the same whether we draw the cards simultaneously or one after the other. The probability will remain 3/56."

If indeed you think so you are wrong. (If not, we apologize.) The point is that we have altered not only the conditions of drawing the cards but also the event itself. In Example 3.3 we required that the letter "c" be the first, "a" the second, and "t" the third. But now the order of the letters is unimportant (it can be "cat" or "act" or "tac" or what). Event A with which we are concerned,

A — the word "cat" can be compiled with the letters drawn out — breaks down into several variants:

A = ("cat" or "act" or "tac" or ...).

How many variants can we have? Apparently, their number is equal to the number of permutations of three elements:

$$P_3 = 1\cdot 2\cdot 3 = 6.$$

We must calculate the probabilities of these six alternatives and add them up according to the probability summation rule. It is easy to show that

these variants are equally probable:

$$P(\text{"act"}) = (2/8)\cdot(3/7)\cdot(3/6) = 3/56,$$

$$P(\text{"tac"}) = (3/8)\cdot(2/7)\cdot(3/6) = 3/56, \text{ etc.}$$

Summing them up, we obtain

$$P(A) = (3/56)\cdot 6 = 9/28.$$

The probability multiplication rule becomes especially simple when events A_1, A_2, ..., A_n are independent*. In this case we must multiply not conditional probabilities but simply the probabilities of events (without any condition):

$$P(A_1 \text{ and } A_2 \text{ and } \ldots \text{ and } A_n) = P(A_1)\ P(A_2)\ \ldots\ P(A_n), \tag{3.9}$$

that is, the probability of the combination of independent events is equal to the product of their probabilities.

EXAMPLE 3.4. A rifleman takes four shots at a target. The hit or miss in every shot does not depend on the result of the previous shots (that is, the shots are mutually independent). The probability of a hit in each shot is 0.3. Find the probability that the first three shots will miss and the fourth will hit.

Solution. Let us denote a hit by "+" and a miss by "−". According to the probability multiplication rule for independent events,

$$P(- \ - \ - \ +) = 0.7\cdot 0.7\cdot 0.7\cdot 0.3 = 0.1029.$$

---

* Several events are called mutually independent if any one of them is independent on any combination of the others. It should be noted that for the events to be mutually independent it is insufficient that they be *pairwise* independent. We can find intricate examples of events that are pairwise independent but are not independent in total. What won't mathematicians come up with next!

And now let us take a somewhat more complicated example.

EXAMPLE 3.5. The same conditions as in Example 3.4. Find the probability of an event A consisting in that out of four shots two (no less and no more) will hit the target.

Solution. Event A can be realized in several variants, for example, $+\ +\ -\ -$ or $-\ -\ +\ +$, and so on. We will not again enumerate all the variants but will simply calculate their number. It equals the number of ways in which two hitting shots can be chosen from four:

$$C(4,\ 2) = \frac{4 \cdot 3}{1 \cdot 2} = 6.$$

This means that event A can be realized in six variants:

$$A = (+\ +\ -\ -\ \text{or}\ -\ -\ +\ +\ \text{or}\ \ldots)$$

(six variants in total). Let us find the probability of each variant. By the probability multiplication rule,

$$P(+\ +\ -\ -) = 0.3 \cdot 0.3 \cdot 0.7 \cdot 0.7 = 0.3^2 \cdot 0.7^2 = 0.0441,$$

$$P(-\ -\ +\ +) = 0.7^2 \cdot 0.3^2 = 0.0441, \text{ etc.}$$

Again the probabilities of all the variants are equal (you shouldn't think that this is always the case!). Adding them up by the probability summation rule, we obtain

$$P(A) = 0.0441 \cdot 6 = 0.2646 \approx 0.265.$$

In connection with this example we shall formulate a general rule which is used in solving problems: if we have to find the probability of an event we must first of all ask ourselves a question: In which way can this

event be realized? That is, we must represent it as a series of mutually exclusive variants. Then we must calculate the probability of each variant and sum up all these probabilities.

In the next example, however, the probabilities of the variants will be unequal.

EXAMPLE 3.6. Three riflemen take one shot each at the same target. The probability of the first rifleman hitting the target is $p_1 = 0.4$, of the second $p_2 = 0.5$, and of the third $p_3 = 0.7$. Find the probability that two riflemen hit the target.

Solution. Event A (two hits at the target) breaks down into $C(3, 2) = C(3, 1) = 3$ variants:

$$A = (+\ +\ -\ \text{or}\ +\ -\ +\ \text{or}\ -\ +\ +),$$

$$P(+\ +\ -) = 0.4 \cdot 0.5 \cdot 0.3 = 0.060,$$

$$P(+\ -\ +) = 0.4 \cdot 0.5 \cdot 0.7 = 0.140,$$

$$P(-\ +\ +) = 0.6 \cdot 0.5 \cdot 0.7 = 0.210.$$

Adding up these probabilities, we obtain

$$P(A) = 0.410.$$

You are probably surprised at the great number of examples with "shots" and "hitting the target". It should be noted that these examples of experiments which cannot be described by the classical model are just as unavoidable and traditional in probability theory as the classical examples with coins and dice for the classical model. And just as the latter do not indicate any special proclivity toward gambling in the persons engaged with probability theory, the examples with shots are not an indication of their bloodthirstiness. As a matter of fact, these examples are the simplest. So be patient and let us have one more example.

EXAMPLE 3.7. The same conditions as in Example 3.4 (four shots with the probability of a hit in each shot

being 0.3). Find the probability that the rifleman hits the target at least once.

Solution. Denote C—at least one hit, and find the probability of event C. First of all let us ask ourselves a question: In what way can this event be realized? We shall see that this event has many variants:

$$C = (+ \ + \ + \ + \text{ or } + \ + \ + \ - \text{ or } \ldots).$$

Of course we could find the probabilities of these variants and add them up. But it is tiresome! It is much easier to concentrate on the opposite event $\overline{C}$—no hit.

This event has only one variant:

$$\overline{C} = (- \ - \ - \ -),$$

and its probability is

$$P(\overline{C}) = 0.7^4 \approx 0.240.$$

Subtracting it from unity, we obtain

$$P(C) \approx 1 - 0.240 = 0.760.$$

In connection with this example we can formulate one more general rule: *If an opposite event divides into a smaller number of variants than the event in question, we should shift to the opposite event.*

One of the indications which allows us to conclude almost for certain that it is worthwhile concentrating on the opposite event is the presence of the words "at least" in the statement of the problem.

EXAMPLE 3.8. Let n persons unknown to each other meet in some place. Find the probability that at least two of them have the same birthday (that is, the same date and the same month).

Solution. We shall proceed on the assumption that all the days of the year are equally probable as birthdays. (In fact, it is not quite so but we can use this

assumption to the first approximation.) Denote the event we are interested in by C—the birthdays of at least two persons coincide.

The words "at least" (or "if only" which is the same) should at once alert us: wouldn't it be better to shift to the contrary event? Indeed, event C is very complex and breaks down into such a tremendous number of variants that the very thought of considering them all makes our flesh creep. As to the opposite event, $\bar{C}$—all the persons have different birthdays, the number of alternatives for it is much more modest, and its probability can be easily calculated. Let us demonstrate this. We represent event $\bar{C}$ as a combination of n events. Let us choose somebody from the people gathered and conditionally call him "the first" (it should be noted that this gives no real advantage to this person). The first could have been born on any day of the year; the probability of this event is one. Choose arbitrarily "the second"—he could have been born on any date but the birthday of the first. The probability of this event is 364/365. The "third" has only 363 days when he is allowed to have been born, and so on. Using the probability multiplication rule, we obtain

$$P(\bar{C}) = 1 \cdot \frac{364}{365} \cdot \frac{363}{365} \cdots \frac{365 - (n-1)}{365}. \quad (3.10)^*$$

from which it is easy to find $P(C) = 1 - P(\bar{C})$.

Note an interesting feature of this problem: with increasing n (even to moderate values) event C becomes practically certain. For instance, already for $n = 50$

---

* This formula is valid only for $n \leqslant 365$; for $n > 365$, obviously, $P(\bar{C}) = 0$. For the sake of simplicity we have ignored leap years and the probability that somebody's birthday is on February 29.

formula (3.10) yields

$$P(\bar{C}) \approx 0.03 \text{ and } P(C) \approx 0.97,$$

that is, event C (with a high confidence level of 0.97) can be considered as practically certain!

This simple calculation may help you (if you wish) to play the role of magician-soothsayer. Suppose at some place many people gather, 50 or a little over (for n much greater than 50 the experiment becomes ineffective), whose birthdays you don't know in advance. You start to assert that among them there are persons whose birthdays coincide. For verification you take a sheet of paper which was divided beforehand into 12 columns (January, February, etc.); each column has 31 rows (the possible number of days in a month) which are labeled on the side. You pass the sheet to the people and suggest that each person put a cross in a box corresponding to his birthday. "As soon as two crosses appear in a box, please return this sheet of paper to me," you declare. "And if not?" somebody asks. You grin self-confidently: "They will appear!" In fact you feel a nagging doubt. You are aware that it is not absolutely certain that your prediction will be fulfilled. There remains. however small, a probability of 0.03 that it will not come true. What if?... Well, you must be ready to run the risk.

Now, moving to more serious things, we shall solve a very important problem which is often encountered in practice in very diversified forms.

PROBLEM 3.1. Let n independent trials be made. In each trial event A occurs with probability p. Find the probability that in n trials event A occurs at least once.

Solution. Denote:

C — at least one occurrence of event A in n trials.

The magic words "at least" lead us to the opposite event. In fact, event $\bar{C}$ (not a single occurrence of event

A) is much simpler than C and has only one variant:

$$\bar{C} = \underbrace{(- \ - \ - \ \ldots \ -)}_{n \text{ times}}.$$

By the probability multiplication rule for independent events we have

$$P(\bar{C}) = (1 - p)^n,$$

whence

$$P(C) = 1 - (1 - p)^n. \tag{3.11}$$

Note general formula (3.11); it is used to solve many practically important problems.

EXAMPLE 3.9. Let the probability of detecting a cosmic object during one scanning cycle of a radar be $p = 0.1$ (the detections in individual cycles are mutually independent). Find the probability that on 10 scanning cycles the object will be detected.

Solution. By formula (3.11) we have

$$P(C) = 1 - (1 - 0.1)^{10} = 1 - 0.9^{10} \approx 1 - 0.348 = 0.652.$$

EXAMPLE 3.10. An installation consists of seven parts (elements). Each part, independently of the others, may have a fault, the probability of this is 0.05. If only one part is faulty, the operation of the installation will break down. Find the probability of this event.

Solution. The probability of event C—"breakdown"—can be calculated by formula (3.11):

$$P(C) = 1 - (1 - 0.05)^7 = 1 - 0.95^7 \approx 1 - 0.695 = 0.305.$$

That's a nice thing! The probability of a breakdown is 0.305, that is, it is greater than 30%! It goes without saying that we have an emergency. We must quickly improve the reliability (the probability of the good working order) of every part!

Note the following circumstance: the state of disrepair of each part has a rather low probability of 0.05. If we inattentively look at this figure we could bear with it and declare the "damage of the part" to be a practically impossible event (as we did several times when we considered the predictions of the results of tossing coins). Now it is quite another kettle of fish! First, we have not one part but seven, and the breakdown of *at least one part* is not a low-probability event. Moreover, the consequences of our inattention (the breakdown!) are not so harmless as an unsuccessful prediction.

It is worth noting that sometimes calculations by probability methods give unexpected results as if "contradicting common sense". We shall illustrate this by a humorous example.

EXAMPLE 3.12. Two hunters — Sam and George — went hunting, saw a bear and simultaneously shot at it. The bear was killed. There was only one hole in its hide. Which of the hunters was responsible was unknown. Though it is more likely that it was Sam — he was older and more experienced, and it was with the probability of 0.8 that he hit the target of such a size and from that distance. George, the younger and less experienced hunter, hit the same target with the probability of 0.4. The hide was sold for 50 rubles. How to divide fairly this sum of money between Sam and George?

Solution. You would probably suggest dividing 50 rubles between Sam and George in proportion to the probabilities 0.8 and 0.4, that is, to give Sam 2/3 of the money (33 rubles and 30 kopeks) and the rest, 16 rubles and 70 kopeks, to George. And, guess what, you would be wrong!

To convince you of this let us somewhat change the conditions of the problem. Suppose Sam hits the bear

on one shot *for sure* (with the probability of 1), and George only with the probability of 0.5. There is only one hole in the bear's hide. Who made it (and hence, who owns the hide)? Of course, it is Sam! Since by the conditions of the problem he couldn't miss. And you would like to divide the money in proportion 2:1, that is, as before, you would give Sam only 2/3 and George 1/3 of the money. Something seems to be wrong in your reasoning. What is it?

The problem is that you divided the money in proportion to the probabilities of a hit on one shot not taking into account that there is only one hole in the bear's hide, which means that one hunter hit and the other missed! This very miss wasn't taken into account.

Let us solve the problem in the proper way. Consider event A — one hole in the hide. How could this event be realized? Clearly, in two ways:

A_1 — Sam hit and George missed,

A_2 — George hit and Sam missed.

The probabilities of these alternatives can be found by the probability multiplication rule:

$$P(A_1) = 0.8 \cdot 0.6 = 0.48,$$

$$P(A_2) = 0.4 \cdot 0.2 = 0.08.$$

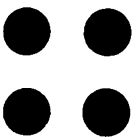
These are the probabilities in proportion to which 50 rubles should be divided. In this case Sam's share will be

$$50 \cdot \frac{0.48}{0.48 + 0.08} = 42.8 \text{ rubles},$$

and for George we have only

$$50 \cdot \frac{0.08}{0.48 + 0.08} = 7.2 \text{ rubles}.$$

Of course it can happen that Sam, after receiving such a lion's share, will invite George for refreshment at the fire as hunters often do.



Random Variables

In this chapter you will become acquainted with the new and very important concept of *random variables.*

Let us briefly recall the contents of the previous chapters. In Chapter 1 you learned how to calculate the probability of an event by the simplest method, that is, by direct calculation of the proportion of favourable outcomes. As soon as you've mastered this method you were disappointed: it turned out that this approach could be used far from universally but only in those (relatively rare) problems where the experiment can be described by the classical model, that is, it possesses a symmetry of possible outcomes. But in the next chapter you learned how to calculate approximately the probability of an event by its frequency; in this case no limitations were placed on the experiment. Just after you became a little familiar with this method you suffered another disappointment: it turned out that this method was not the principal one in probability theory! In Chapter 3 (Basic Rules of Probability Theory) you at last approached the methods which, as follows from the title, are fundamental in

this field of science. "At last!" you think. "We have managed to reach the very kernel, the essence! We will be in for no more disappointments now!" Alas! You are to meet one more (this time the last) disappointment. As a matter of fact, the apparatus of events we have dealt with so far is not the principal one in modern probability theory!

"Then which apparatus is the principal one?" you may ask, boiling with anger. "By George, why did you make me learn not the principal apparatus?"

"The trouble is," we shall answer patiently, "that without knowledge about all this it is impossible even to start studying the modern apparatus."

"And what is this apparatus?"

"It is the apparatus of random variables."

This chapter will be devoted to the basic concepts of modern probability theory—random variables, different kinds of random variables, and methods of describing and manipulating them. We shall not consider random variables in detail as we did with random events but present them "descriptively" since in this case, should we try to use an appropriate mathematical apparatus, it would be more complicated and could alienate the beginner. And we want, on the contrary, to make the beginner's "first steps" easier. Thus, let us get acquainted with the concept of random variable.

A quantity is called a random variable if as a result of an experiment it may take on various values which are unknown beforehand.

As is always the case in probability theory, this definition is rather vague and contains some indeterminacy, uncertainty.... To comprehend the notion of a random variable you must simply get used to it. For this purpose consider examples of random variables.

1. The experiment consists in tossing two coins. The number of heads appearing as a result is a random variable. Its possible values are 0, 1, and 2. Which of them it will take is not known beforehand.

2. A student is going to sit for his exam. The mark he will get is a random variable. Its possible values are 2, 3, 4, and 5.

3. There are 28 students in a group. On any one day the number of absent students is registered. This is a random variable. Its possible values are 0, 1, ..., 28.

"Go on!" you may say." All 28 students cannot get ill (or be absent) at once."

"Yes, it is a practically impossible event. But who said that all the values of the random variable should be equally likely?"

All random variables in these examples belong to the type of so-called *discrete random variables.* The random variable is called discrete if its possible values are separated from one another by some interval. On the x-axis these values are depicted by *separate points.*

There are random variables of the other kind, which are called *continuous.* The values of these random variables continuously fill a certain interval on the number axis. The boundaries of the interval are sometimes sharp and sometimes blurred and undetermined. Let us consider several examples of continuous random variables.

1. The operating time between two failures (errors) of a computer. The values of this random variable continuously fill a certain interval of the number axis. The lower boundary of this interval is quite definite (0) and the upper boundary is blurred, undetermined, and can be found only as a result of experiments.

2. The weight of a freight train arriving at a station for unloading.

3. The height of rising water during a spate.

4. The error appearing when a body is weighed by an analytical balance (as distinct from the previous problems, this random variable can take on both positive and negative values).

5. The density of milk taken for testing at a farm.

6. The time spent by an eight-former in front of the television set.

It should be emphasized that to speak of a random variable in the sense used in probability theory it is necessary to *define the experiment* in which the random variable takes on this or that value. For instance, in the first example concerning continuous random variables it should be pointed out what kind of a computer we are talking about, what is its age and operating conditions. In the fourth example we must clarify which kind of a balance we are using for weighing and which set of small weights is used. The necessity of clarifying the experimental conditions should be always borne in mind though for the sake of brevity we shall not always specify the details.

Note the following circumstance. In fact all random variables which we call continuous can be measured in certain units (minutes, centimetres, tons, etc.), and in this sense, strictly speaking, they are discrete. For example, it is senseless to measure the random variable "the height of a man" more accurately than to within 1 cm, so we obtain essentially the discrete random variable with the values separated by the interval of 1 cm. But the number of these values is very large and they are situated on the number axis very "densely". So in this case it is more convenient to treat this random variable as a continuous one.

From now on let us denote random variables by Latin capitals and their possible values by the respective lower cases (for example, a random variable is X and its possible values are x_1, x_2, ...). The term

random variable we shall sometimes abbreviate to RV.

So, suppose we have RV X with some of its values. Naturally, not all these values are equally likely. There are among them some more probable and some less probable. We shall call the *distribution function* (law) of the random variable any function that describes the distribution of the probabilities among the values of this variable. We shall show you not all the forms of distributions but only some of them, the simplest ones.

The distribution of a discrete random variable can be represented in the form of the so-called *distribution series*. This is the name given to a table with two rows; the upper row enumerates all possible values of the random variable: x_1, x_2, ..., x_n, and the lower row gives the probabilities corresponding to them: p_1, p_2, ..., p_n:

x_i	x_1	x_2	...	x_n
p_i	p_1	p_2	...	p_n

Each probability p_i is nothing other than the probability of the event consisting in that RV X takes on the value x_i:

$$P_i = P(X = x_i) \ (i = 1, \ldots, n).$$

The sum of all probabilities p_i is obviously equal to one:

$$p_1 + p_2 + \ldots + p_n = 1.$$

The unity is somehow distributed over the values of RV, hence the term distribution.

EXAMPLE 4.1. Three independent shots are taken at a target; the probability of a hit each time is $p = 0.4$. Discrete RV X is the number of hits. Construct its distribution series.

Solution. In the upper row write down all the possible values of RV X : 0, 1, 2, and 3, and in the lower row, the respective probabilities which we denote by p_0, p_1, p_2, and p_3. We already know how to calculate them (see Chapter 3). We have:

$$p_0 = P(- \ - \ -) = 0.6^3 = 0.216,$$

$$p_1 = P(+ \ - \ - \quad \text{or} \quad - \ - \ + \quad \text{or} \quad - \ + \ -) = 3 \cdot 0.6^2 \cdot 0.4 = 0.432,$$

$$p_2 = P(+ \ + \ - \quad \text{or} \quad + \ - \ + \quad \text{or} \quad - \ + \ +) = 3 \cdot 0.4^2 \cdot 0.6 = 0.288,$$

$$p_3 = P(+ \ + \ +) = 0.4^3 = 0.064.$$

Let us check whether or not the sum of these probabilities is equal to unity. It really is, so that's all right. The distribution series for RV X has the form

x_i	0	1	2	3
p_i	0.216	0.432	0.288	0.064

EXAMPLE 4.2. A sportsman makes several attempts to throw a ball into a basket. At every attempt (independently of the others) a success occurs with the probability p. As soon as the ball gets into the basket the exercise is stopped. Discrete RV X is the number of trials to be made. Construct the distribution series for RV X.

Solution. Possible values of RV X are: $x_1 = 1$, $x_2 = 2$, ..., $x_k = k$, and so on, theoretically to infinity. Find the probabilities of all these values: p_1 is the probability that the sportsman pockets the ball on the first attempt. Obviously, it is equal to p, $p_1 = p$. Let us find p_2, the probability that two attempts will be made. For this the combination of two events must occur: (1) on the first attempt the sportsman missed the basket

and (2) on the second attempt he hit it. The probability $p_2 = (1 - p)p$. Similarly we find $p_3 = (1 - p)^2 p$ (the first two attempts were unsuccessful but the third attempt was lucky), and in general, $p_i = (1 - p)^{i-1} p$. The distribution series for RV X has the form:

x_i	1	2	3	...	i	...
p_i	p	$(1 - p)p$	$(1 - p)^2 p$	...	$(1 - p)^{i-1} p$	...

It should be noted that probabilities p_i form a geometric progression with the common ratio $(1 - p)$; for this reason such a probability distribution is called the geometric distribution.

Now let us see how we can determine the probability distribution for a continuous RV. We cannot construct a distribution series for a random variable which continuously fills a certain interval on the x-axis.

"Why not?" you may ask. Here is the answer. To begin with, we cannot write even the upper row of the series, that is, enumerate one after the other all the possible values of RV. Indeed, whatever pair of values we take, there can always be found some values between them. (For those who are familiar with set theory we may add that the number of these values is *uncountable*). Moreover, we shall be faced with another difficulty: should we try to assign a certain probability to each individual value of a continuous RV, we would find that this probability equals ... zero! Yes, exactly zero. It isn't a slip of the pen, and you'll see it for yourself.

Imagine that we are on a beach covered with pebbles. We are interested in a random variable X — the weight of a separate stone.

Well, let's weigh the stones. First we shall start with a reasonable accuracy to within 1 g. We shall assume

that the weight of a stone is 30 g if it is equal to 30 g to within 1 g. We shall obtain some frequency of the weight 30 g—we don't know now what it is equal to and this is not important.

Now let us increase the accuracy and weigh the stones to within 0.1 g, that is we shall consider the weight of a stone to be of 30 g if it is equal to 30 g to within 0.1 g. In this case some stones which we considered to weigh 30 g at the first rough weighing will be excluded. The frequency of the event $X = 30$ g will decrease. How many times? Approximately ten times.

And now let us increase the accuracy still further and weigh the stones to within... 1 mg! The frequency of appearance of the weight 30 g will become a further 100 times lower.

And the frequency, which is a close relative of the probability, approaches it as the number of experiments increases (there are many stones on the beach, so don't worry—we'll always have enough specimens for weighing!). So, what probability should be ascribed to an event consisting in the fact that the weight of a stone is exactly 30 g, no more and no less? Clearly, zero probability—there is nothing to be done!

You are astonished. Maybe you are even indignant. As you remember, zero probability applies to *impossible events*. But the event consisting in the fact that a continuous RV X takes on a certain value x is possible! How can it be that its probability is zero?

Let us recollect. Indeed, we assertained that the *probability of an impossible event* is zero, and it is true. But have we stated that any event with zero probability is impossible? Not at all! Now we have come across possible events whose probabilities are zero.

Don't be in a hurry. Let us think a little. Forget for

a moment probability theory and just imagine a plane figure of area S. And now take any point within this figure. What is the area of this point? Obviously, it is zero. The figure undoubtedly consists of points each of which has zero area, and the whole figure has a nonzero area. This "paradox" does not astonish you—you simply get accustomed to it. In the same manner you should get used to the fact that the probability of getting at every individual point for a continuous random variable is exactly zero.*

"Then how can we speak of the *probability distribution* for a continuous random variable?" you may ask. "Each of its values has one and the same zero probability, hasn't it?"

You are quite right. There is no sense in speaking of the probability distribution of a continuous RV over its *separate values.* And still the distribution for such a random variable exists. For example, there is no doubt that the value of man's height of 170 cm is more probable than 210 cm, though both values are possible.

Here we must introduce one more important concept—*probability density.*

The notion "density" is familiar to you from physics. For example, the density of a substance is its mass per unit volume. And if the substance is inhomogeneous? Then we have to consider its *local* density. In probability theory we shall also consider local density (that is, the probability at point x per unit length).

Probability density of a continuous random variable

---

* We should stipulate that there are random variables of a special, so-called mixed type. In addition to a continuous interval of possible values with zero probabilities they also have separate singular values with nonzero probabilities. We shall not consider these intricate random variables but you should know that they exist.

X is the limit of the ratio of the probability of getting RV *X into a small interval in the vicinity of point x to the length of this interval as the latter tends to zero.*

Probability density can be easily derived from frequency density which is related to it. Consider a continuous RV X (for example, the height of a man or the weight of a stone on the beach). First of all let us conduct a number of experiments with this random variable, in each of which it takes on a certain value (for example, we measure the height of a group of people or weigh many stones). We are interested in the distribution of probabilities for our random variable. Well, let us divide the whole range of values of RV X into certain intervals (or, as they say , categories), for example, 150-155 cm, 155-160 cm, 160-165 cm, ..., 190-195 cm, 195-200 cm. Calculate how many values of RV X are put into each category * and divide by the total number of experiments made. We obtain the "frequency of the category" (clearly, the sum of the frequencies of all categories should be equal to one). Now calculate the *frequency density* for each interval; for this purpose divide the frequency by the length of the interval (generally the lengths of the intervals can be different).

If we have a sufficiently large data file (say, of the order of several hundreds or more, which is better), then plotting for our RV the frequency density, we can obtain a clear idea about its probability density. When treating these statistical data it is very convenient to plot a kind of a graph called a histogram (Fig. 2). The histogram is plotted as follows. On each interval, as on the base, we plot a rectangle whose area is equal to the interval frequency (and hence, its height is equal to the

---

* If some value of X exactly coincides with the boundary of two intervals, we give half of it to each interval.

frequency density). The area bounded by the histogram is obviously equal to one. With an increasing number N of experiments the intervals become shorter and shorter and the step form of the histogram becomes smoother, approaching a certain smooth curve which is called the *distribution curve* (Fig. 3). Along the y-axis

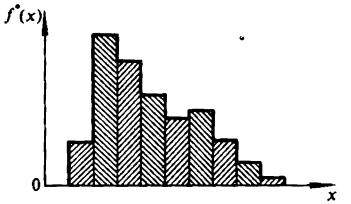


Fig. 2

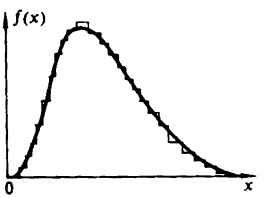


Fig. 3

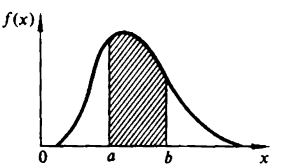


Fig. 4

there will be not the frequency density but the probability density. Clearly, the area bounded by the distribution curve equals unity just as the area of its "relative", the hystogram. The probability that RV X lies within the interval (a, b) equals the area resting on this interval (Fig. 4). If we denote the probability density function by $f(x)$, the probability that RV X fits into interval (a, b) will be expressed by the definite integral:

$$P(a, b) = \int_a^b f(x)\,dx. \tag{4.1}$$

Thus if we spare no time and effort, the probability density function $f(x)$ we are interested in can be determined from the experimental data with an arbitrary accuracy. But is it worth powder and shot? Do we really need an *absolutely accurate value* of probability density $f(x)$? No, very often we don't, it is sufficient to have an approximate idea of the distribution of RV X (all probability calculations are approximate, rough estimates by nature). And to know approximately the distribution law for RV X we needn't make a tremendous number of experiments but use a moderate number of 300-400 experiments and sometimes even less. Plotting the hystogram according to the available experimental data we can then level it out using some smooth curve (of course, it must be a curve which bounds a unit area). Probability theory has at its disposal a great number of curves satisfying this condition. Some of them have certain advantages with respect to others since their integral (4.1) can be easily evaluated or there are tables of the values compiled for this integral. In other cases the conditions of the appearance of a random variable suggest a certain type of distribution resulting from theoretical considerations. We shall not go into such a detail here — this is a special question. It should be emphasized, however, that in finding the distribution of random variables of principal importance there are not *direct* but *indirect* methods which enable us to find the distributions of some random variables not directly from the experiment but through available data on other kinds of random variables related to the first ones.

In the realization of these indirect methods an important part is played by the so-called *numerical characteristics* or random variables.

Numerical characteristics are certain numbers that characterize certain properties of random variables and

their characteristic features, such as the mean value around which a random spread takes place, the degree of this spread (so to speak, the degree of randomization of a random variable) and a number of other characters. It turns out that many problems in probability theory can be solved without (or almost without) using distribution laws but using only numerical characteristics. Here we give you only two (but the most important) numerical characteristics of random variables — *expectation* and *variance.*

Expectation (or expected value) $E[X]$ of a discrete random variable X is the sum of the products of all its possible values by the respective probabilities:

$$E[X] = x_1 p_1 + x_2 p_2 + \ldots + x_n p_n \qquad (4.2)$$

or, using a summation sign,

$$E[X] = \sum_{i=1}^{n} x_i p_i, \qquad (4.3)$$

where $x_1, x_2, \ldots, x_n$ are possible values of RV X and p_i is the probability that RV X takes on the value x_i.

As can be seen from formula (4.3), the expectation of random variable X is nothing else but the "mean weighted" value of all its possible values, and each of these values enters the sum with the "weight" equal to its probability. If there is an infinite number of possible values of RV (as in Example 4.2), then the sum (4.3) consists of an infinite number of addends.

The expectation or the mean of RV is, as it were, its representative which can replace it in rough estimates. Essentially, this is always the case when we ignore randomness in some problems.

EXAMPLE 4.3. Find the expectation of RV X considered in example 4.1 (the number of hits out of three shots).

Solution. By formula (4.3) we have

$$E[X] = 0 \cdot 0.216 + 1 \cdot 0.432 + 2 \cdot 0.288 + 3 \cdot 0.064 = 1.2.$$

For a continuous RV X we can also introduce the concept of expectation; naturally, the sum in formula (4.3) should be replaced by the integral:

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx, \tag{4.4}$$

where $f(x)$ is the probability density function for continuous RV X.

Let us now discuss expectation, its physical meaning, and "genealogy". Just as probability has a close relative—frequency, expectation also has a close relative—arithmetic mean of the results of observations. Just as frequency approaches probability with increasing number of experiments, in the same way arithmetic mean of observed values of RV approaches its expectation as the number of experiments increases.

Let us prove this for discrete random variables (we hope that you will readily take our words on trust that the same is true for continuous random variables). Thus, suppose we have a discrete RV X with the distribution series:

x_i	x_1	x_2	...	x_n
p_i	p_1	p_2	...	p_n

Suppose we conducted N experiments as a result of which the value x_1 appeared M_1 times, x_2 appeared M_2 times, and so on. Find the arithmetic mean of the observed values of RV X (we denote it by $\bar{X}$):

$$\bar{X} = \frac{x_1M_1 + x_2M_2 + \ldots + x_nM_n}{N} = \sum_{i=1}^{n} x_i \frac{M_i}{N}.$$

But $\frac{M_i}{N}$ is exactly the frequency of the event $(X = x_i)$; denote it by p_i^*:

$$\frac{M_i}{N} = p_i^*,$$

whence

$$\overline{X} = \sum_{i=1}^{n} x_i p_i^*.$$

We know that as the number N of experiments increases, the frequency p_i^* of the event with practical certainty approaches its probability p_i, hence *the arithmetic mean of the observed values of a random variable with increasing number of experiments with a practical certainty will be arbitrarily close to the expectation.*

This statement is one of the forms of the *law of large numbers* (the so-called Chebyshev theorem), which plays an important role in practical applications of probability theory. In fact, just as unknown probability of the event can be approximately determined by its frequency in the long series of experiments, the expectation of RV X can be approximately found as the arithmetic mean of its observed values:

$$E[X] \approx \overline{X}. \tag{4.5}$$

It should be noted that in order to calculate the expectation of a random variable in which we are interested from the results of experiments, it is not necessary to know its distribution; it is sufficient just to calculate the mean of all the results of observations.

One more remark: to find the expectation of a random variable with satisfactory accuracy, we need

not make the same number of experiments (of the order of several hundreds) as for plotting a hystogram; it suffices to make a much smaller number of experiments (of the order of dozens).

Now let us introduce another very important numerical characteristic of a random variable — its *variance*. The variance means the spread of the values of the random variable around its mean. The greater the variance, the more "chancy" is the random variable.

The variance of RV X is calculated as follows: the expectation (the mean) is subtracted from each possible value of the random variable. The obtained deviations from the mean are squared, multiplied by the probability of the respective values, and all these products are summed up. The result is the variance which we denote by $D[X]$ or σ^2:

$$\sigma^2 = D[X] = (x_1 - E[X])^2 p_1 + (x_2 - E[X])^2 p_2 + \ldots + (x_n - E[X])^2 p_n,$$

or, in a short form,

$$\sigma^2 = D[X] = \sum_{i=1}^{n} (x_i - E[X])^2 p_i. \qquad (4.6)$$

The question may arise: why do we square the deviations from the mean? It is done to get rid of the sign (plus or minus). Of course, we could get rid of the sign just by dropping it (taking the deviation modulo), but in that case we would obtain a much less convenient characteristic than the variance.

EXAMPLE 4.4. Find the variance of the number of hits X in Example 4.1.

Solution. From formula (4.6) we have

$$D[X] = (0 - 1.2)^2 \cdot 0.216 + (1 - 1.2)^2 \cdot 0.432 + (2 - 1.2)^2 \cdot 0.288 + (3 - 1.2)^2 \cdot 0.064 = 0.72.$$

It should be noted that formula (4.6) is not the most convenient one for calculating the variance. The variance can be (and usually is) calculated by the formula

$$D[X] = E[X^2] - (E[X])^2, \tag{4.7}$$

that is, *the variance of a random variable X is the expectation of the square of X minus the square of the expectation of X*.

This formula can be readily derived from (4.6) using identity transformations, but we shall not stop to do this. Instead we shall verify the validity of this formula using the preceding example:

$$D[X] = 0^2 \cdot 0.216 + 1^2 \cdot 0.432 + 2^2 \cdot 0.288 + 3^2 \cdot 0.064 - (1.2)^2 = 0.72.$$

For continuous random variables the variance is calculated by the formula similar to (4.6), but naturally summation is replaced by integration:

$$D[X] = \int_{-\infty}^{\infty} (x - E[X])^2 f(x)\,dx. \tag{4.8}$$

Usually it is more convenient to use the formula analogous to (4.7), which in this case reads

$$D[X] = \int_{-\infty}^{\infty} x^2 f(x)\,dx - (E[X])^2. \tag{4.9}$$

Just as it is not necessary to know the distribution law to determine the expectation, the variance can also be approximately calculated directly from the results of experiments by averaging the squares of deviations of the observed values of RV from their arithmetic mean:

$$D[X] \approx \frac{1}{N} \sum_{k=1}^{N} (x_k - \overline{X})^2, \tag{4.10}$$

where k is the serial number of an experiment, x_k is the value of RV X observed in the k-th experiment, and N is the total number of experiments.

And again it is more convenient to calculate the variance by the formula "mean square minus squared mean":

$$D[X] \approx \frac{1}{N} \sum_{k=1}^{N} x_k^2 - \overline{X}^2. \tag{4.11}$$

Formulas (4.10) and (4.11) can be used only for rough estimates of the variance even with a not very large number of experiments (better something than nothing!); naturally, they illustrate the variance with a not very high accuracy. In mathematical statistics in such cases it is used to introduce "a correction due to small number of experiments" and multiply the obtained result by the correction factor $N/(N-1)$. This correction should not be overestimated. You should remember that with a very small number of experiments nothing very useful can be obtained from their statistical treatment (however hard you try!), and with large N the correction factor is close to unity.

The variance of a random variable, being the characteristic of a spread, has a very inconvenient feature: its dimension, as can be seen from formula (4.6), is equal to squared dimension of RV X. For instance, if RV X is expressed in minutes, its variance is measured in "square minutes", which looks meaningless. For this reason the square root is extracted from the variance; we obtain a new characteristic of the spread — the so-called standard deviation:

$$\sigma_X = \sqrt{D[X]}. \tag{4.12}$$

Standard deviation is a very visual and convenient

characteristic of a spread. It readily gives the idea of the amplitude of oscillations of RV about the mean value. For most practically encountered random variables *with a practical certainty we can assert that they deviate from their expectations by not more than* $3\sigma_X$. The confidence level depends on the distribution of RV; for all practical cases it is rather high. The above rule is called the "three sigma rule".

Thus, if we managed to find by some method the two numerical characteristics of RV X — its expectation and standard deviation, we immediately have a tentative idea about the limits of its possible values.

Here you may ask a question: If we discovered these characteristics from experiment, who on earth can prevent us from determining the limits of possible values from the very same experiment?

Well, you are quite right in the case when these characteristics have been found directly from the experiments. But in probability theory these (direct) methods of determining numerical characteristics are not the principal ones. We again must emphasize that not direct but indirect methods are basic methods, which make it possible to find the numerical characteristics of random variables we are interested in, by the numerical characteristics of other random variables related to them. In such cases *basic rules* of arithmetic are applied to numerical characteristics. Some of these rules we give below (naturally without proof).

1. *Additivity of expectation.* The expectation of the sum of random variables is equal to the sum of expectations of addends.

2. *Additivity of variance.* The variance of the sum of independent random variables is equal to the sum of the variances of addends.

3. *Taking a constant factor outside the expectation:* $E[cX] = cE[X]$.

4. *Taking a constant factor outside the variance.* The squared constant factor c can be taken outside the variance:

$$D[cX] = c^2 D[X].$$

All these rules look quite natural except, perhaps, the last one. To convince yourself of its validity, consider the following example. Suppose we doubled a random variable X. Its expectation is naturally also doubled. The deviation of a separate value from the mean also becomes twice as large while its square increases four-fold!

As you will see now, even this small set of rules is sufficient to solve some interesting problems.

PROBLEM 4.1. Let N independent experiments be made; in each of them an event A occurs with probability p. We consider a random variable—the number of experiments in which event A occurs (in short, the number of occurrences of event A). Find the expectation and the variance of X.

Solution. Represent X as the sum of N random variables:

$$X = X_1 + X_2 + \ldots + X_N = \sum_{k=1}^{N} X_k, \qquad (4.13)$$

where RV X_k equals unity if in the k-th experiment event A has occurred and equals zero if event A hasn't occurred. Now use the additivity of expectation:

$$E[X] = \sum_{k=1}^{N} E[X_k]. \qquad (4.14)$$

As the experiments are independent, random variables $X_1, X_2, \ldots, X_N$ are also independent. The additivity of variance yields

$$D[X] = \sum_{k=1}^{N} D[X_k]. \tag{4.15}$$

Now let us find the expectation and the variance of each of the random variables X_k. Take any one of them. Each RV is discrete and has two possible values, 0 and 1, with the probabilities $(1 - p)$ and p. The expectation of such a random variable is

$$E[X_k] = 0 \cdot (1 - p) + 1 \cdot p = p.$$

The variance can be found by formula (4.7):

$$D[X_k] = 0^2 \cdot (1 - p) + 1^2 \cdot p - (E[X])^2 = p - p^2 = p(1 - p).$$

Substituting this into (4.14) and (4.15), we obtain the required quantities

$$E[X] = Np, \; D[X] = Np(1 - p).$$

PROBLEM 4.2. Let N independent trials be made, in each of which event A occurs with probability p. We consider a random variable P^*, the frequency of event A in this series of trials. Find approximately the range of possible values of RV P^*.

Solution. By definition, the frequency is the ratio of the number X of occurrences of A to the total number of trials N:

$$P^* = \frac{X}{N}.$$

Let us find numerical characteristics (that is the ex-

pectation and the variance) of this random variable. Using properties 3 and 4, we obtain

$$E[P^*] = E\left[\frac{X}{N}\right] = \frac{1}{N}\ E[X] = \frac{Np}{N} = p,$$

$$D[P^*] = D\left[\frac{X}{N}\right] = \frac{1}{N^2}\ D[X]$$

$$= \frac{Np(1-p)}{N^2} = \frac{p(1-p)}{N}.$$

Taking the square root of the variance, we find the standard deviation σ_{P^*}:

$$\sigma_{P^*} = \sqrt{\frac{p(1-p)}{N}}.$$

And now use the three sigma rule to find approximately the range of practically possible values of RV P^*:

$$p \pm 3\sqrt{\frac{p(1-p)}{N}}.$$

"By George, it is our old acquaintance!" you would exclaim, if you were attentive. "You gave us this very formula for the confidence interval into which the value of the frequency of an event would fit with the probability of 0.997 when the number N of trials is large. Along with this formula (and even preferably) you recommended another formula with 2 and not 3 in front of the square root. If we remember correctly, that formula held with the probability 0.95. So we have formulas now. But where did the probability come from?"

Wait a little, be patient. To understand from where we obtained the probabilities of 0.997 and 0.95

you should become acquainted with a very important distribution — the so-called *normal distribution.*

Consider a continuous random variable X. It is said to have normal distribution if its probability density function is expressed by the formula

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}, \qquad (4.16)$$

where π is the number familiar to you from geometry and e is the base of natural logarithms ($e \approx 2.71828...$). The normal distribution curve has a symmetric bell-shaped form (Fig. 5) and reaches the maximum at

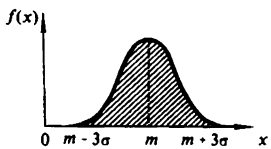


Fig. 5

point m. As we move further from m, the density function drops and in the limit tends to zero. Formula (4.16) contains two parameters: m and σ. As you have probably guessed, m is exactly the expectation and σ is the standard deviation. With the variation of m the curve will move to and fro along the x-axis. If we vary σ, the curve $f(x)$ will change its shape becoming "flatter" with increasing σ and becoming needle-shaped as σ decreases.

A special part played by the normal distribution in probability theory is due to its remarkable property. It turns out that if we add up a large number of independent (or weakly dependent) random variables comparable by the order of variances, then the *distribution of the sum will be close to the normal distribution,*

irrespective of the distributions of the addends, and the closer the greater number of random variables are added up. The above statement is a rough statement of the so-called central-limit theorem which plays a very important role in probability theory. This theorem has a great number of modifications differing in the conditions which must be satisfied by random variables in order that their sum be normalized with the increasing number of addends.

In practice many random variables are formed by the "summation principle" and hence have normal or almost normal distribution. For instance, such are the errors in various measurements which are the sums of many "elementary" and practically independent errors due to different causes. As a rule, the errors in shooting, guidance and matching (registration) have normal distribution. The deviations of voltage in a circuit from the nominal value are also due to the resultant action of many independent causes which are added up. Such random variables as the total payment to an insurance company over a long period or the idle time of a computer per year have normal (or close to normal) distributions. Let us show, in particular, that with a large number N of trials such an interesting random variable as the *frequency of an event* also has approximately normal distribution. In fact,

$$P^* = \frac{X_1 + X_2 + \ldots + X_n}{N} = \sum_{k=1}^{N} \frac{X_k}{N},$$

where X_k is a random variable equal to one if in the k-th trial event A occurs, and equal to zero if it doesn't. It is clear that for a large number N of trials frequency P^* is the sum of a great number of independent

addends, each of them having the same variance

$$D\left[\frac{X_k}{N}\right] = \frac{1}{N^2}\, p(1 - p).$$

Hence we may conclude that frequency P^* of event A has a normal distribution when the number N of trials is large.

As the normal distribution is frequently encountered in practice, it is often necessary to calculate the probability that RV X, having a normal distribution fits into the limits of interval (a, b). Integral (4.1) cannot be expressed in the case in terms of elementary functions; for its calculation use is made of the tables compiled for a special function — the so-called *Laplace function*

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2}\, dt.$$

Here we give an excerpt of the tables for the Laplace function.

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0.0	0.0000	1.0	0.3413	2.0	0.4772	3.0	0.4986
0.1	0.0398	1.1	0.3643	2.1	0.4821	3.1	0.4990
0.2	0.0793	1.2	0.3849	2.2	0.4861	3.2	0.4993
0.3	0.1179	1.3	0.4032	2.3	0.4893	3.3	0.4995
0.4	0.1554	1.4	0.4192	2.4	0.4918	3.4	0.4997
0.5	0.1915	1.5	0.4332	2.5	0.4938	3.5	0.4998
0.6	0.2257	1.6	0.4452	2.6	0.4953	3.6	0.4998
0.7	0.2580	1.7	0.4554	2.7	0.4965	3.7	0.4999
0.8	0.2881	1.8	0.4641	2.8	0.4947	3.8	0.4999
0.9	0.3159	1.9	0.4713	2.9	0.4981	3.9	0.5000

For $x \geqslant 4$ we can take $\Phi(x) = 0.5000$ to within four decimal places.

When using this table it should be borne in mind that the Laplace function is odd, that is, $\Phi(-x) = -\Phi(x)$.

The probability that RV X having a normal distribution with parameters m and σ fits into the limits of interval (a, b) is expressed in terms of the Laplace function by the formula

$$P(a, b) = \Phi\left(\frac{b - m}{\sigma}\right) - \Phi\left(\frac{a - m}{\sigma}\right). \quad (4.17)$$

EXAMPLE 4.5. Find the probability that RV X having a normal distribution with parameters m and σ (a) deviates from its mean by less than 2σ; (b) deviates from its mean by less than 3σ.

Solution. Using formula (4.17) and the table, we find

$$\text{(a) } P(m - 2\sigma, m + 2\sigma) = \Phi\left(\frac{m + 2\sigma - m}{\sigma}\right) - \Phi\left(\frac{m - 2\sigma - m}{\sigma}\right)$$

$$= \Phi(2) - \Phi(-2) = \Phi(2) + \Phi(2) = 0.4772 + 0.4772 = 0.9544 \approx 0.95.$$

$$\text{(b) } P(m - 3\sigma, m + 3\sigma) = \Phi(3) - \Phi(-3)$$

$$= \Phi(3) + \Phi(3) = 0.4986 + 0.4986 = 0.9972 \approx 0.997.$$

At last we have these two confidence coefficients: 0.95 and 0.997 with which we set confidence intervals for the frequency of an event in Chapter 2. It was a difficult struggle but we got there!

But now, having the normal distribution at hand, we can solve some instructive problems.

EXAMPLE 4.6. A train consists of 100 waggons. The

weight of each waggon is a random variable with expectation $m_q = 65$ tons and standard deviation $\sigma_q = 9$ tons. The locomotive can pull the train if its weight is not more than 6600 tons, otherwise an auxiliary locomotive is needed. Find the probability that there will be no need of an auxiliary locomotive.

Solution. The weight X of the train can be represented as the sum of 100 random variables Q_k—the weights of individual cars:

$$X = \sum_{k=1}^{100} Q_k.$$

These random variables have the same expectations $m_q = 65$ and the same variance $D_q = \sigma_q^2 = 81$. By the summation rule for expectations we have

$$E[X] = 100 \cdot 65 = 6500.$$

By the variance summation rule we obtain

$$D[X] = 100 \cdot 81 = 8100.$$

Taking the square root of $D[X]$, we find the standard deviation

$$\sigma_X = \sqrt{D[X]} = 90.$$

For one locomotive to be able to pull the train the weight of the train must be admissible, that is, it must fit into the interval (0, 6600). The random variable X, the sum of 100 addends, can be considered to have the normal distribution. By formula (4.17) we obtain

$$P(0,\ 6600) = \Phi\left(\frac{6600 - 6500}{90}\right) - \Phi\left(\frac{0 - 6500}{90}\right)$$

$$= \Phi\left(\frac{100}{90}\right) - \Phi\left(\frac{-6500}{90}\right) \approx \Phi(1.1) - \Phi(-72)$$

$$= \Phi(1.1) + \Phi(72) \approx 0.387 + 0.500 = 0.887.$$

So the locomotive can "handle" the train with the probability of 0.887.

Now let us decrease the number of waggons by 2, that is, take $N = 98$. Try to find for yourself the probability that the locomotive can pull such a train. I think that the result of the calculation will surprise you: the probability is approximately 0.99, that is, the event in question is practically sure! And for this we just had to leave two waggons...

Now you see what interesting problems we can solve dealing with combinations of a large number of random variables.

Naturally, here a question arises: How large is "large"? How many random variables must we take for their sum to have the normal distribution? It depends on the distributions of the addends. There are such intricate distributions that normalization can be achieved only with a rather large number of addends. We may say again: How smart the mathematicians are! But nature does not create such inconveniences on purpose. In order to use the normal distribution in practice it is usually enough to have 5-6 or 10 or 20 addends at most (particularly if they all have the same distributions).

How quickly the distribution of the sum of random variables having similar distributions becomes normal can be shown by an example. (You must again take it for granted — nothing else can be done! We have not yet deceived you.) Suppose we have a continuous RV with a constant density distribution over the interval (0, 1) (Fig. 6). The distribution curve is of a rectangular type. How far from the normal distribution it seems to be! But if we sum up two such (independent) random variables, we shall obtain a new RV with the so-called Simpson distribution (Fig. 7), which differs from the normal distribution, but now less so... . If we sum up

three such random variables with constant distributions, we shall obtain a RV with the distribution curve shown in Fig. 8. This curve consists of three sections of parabolas and highly resembles the normal distribution. And if we sum up six random variables with constant distributions, we shall obtain

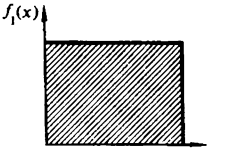


Fig. 6

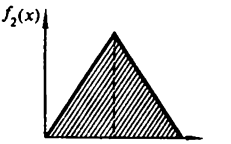


Fig. 7

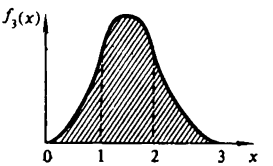


Fig. 8

a curve which cannot be distinguished from the normal distribution at all. This fact justifies the widely used method of obtaining the normal distribution of random variables in computer modelling of random phenomena: it is sufficient to sum up six independent random variables with uniform distributions over the interval (0, 1). Incidentally, most types of computers are equipped with generators of such random variables.

And still we must not become keen over the normal distribution and announce that the distribution of the sum of several random variables is normal. At first we must see (at least to the first approximation) what distributions they have. For example, if they are strongly asymmetric, we shall need a large number of addends.

In particular, the rule that "with a large number of trials the frequency has normal distribution" is to be used cautiously. If the probability p of an event in one of the trials is very low (or, on the contrary, close to one), in this case we shall need an enormous number of trials. By the way, there is a practical method which enables us to verify whether or not the normal distribution can be used for the frequency of the event. We must construct the confidence interval for the frequency of the event using the known method (with a confidence level of 0.997):

$$p \pm 3\sqrt{\frac{p(1-p)}{N}},$$

and if the whole interval (its both ends) does not leave reasonable limits for frequency and hence, for probability (from 0 to 1), we can use the normal distribution. If, however, one of the boundaries of the interval turns out to be outside the interval (0, 1), then the normal distribution cannot be used. In this case for the approximate solution of the problem we must use another distribution—the so-called Poisson distribution. But this distribution (as well as a number of other distributions, of which there are very many in probability theory) is beyond the score of this book.

Still we think that after reading this simple book you will have gained some idea about probability theory and its problems. Possibly, you couldn't stand it, in which case the first acquaintance will also be the last for you. It's a pity, but there's nothing to be done: there are people (even among mathematicians) who loathe probability theory.

But possibly (and that was the purpose of this book) you found the concepts, methods, and possibilities of probability theory interesting. Then you

may study it in greater detail (see Literature). We make no secret of the fact that a deep understanding of probability theory will require considerable mental effort, much more serious than those you made in these "first steps". The next steps will be more difficult but more interesting. You must know the saying: "Learning has bitter roots but sweet fruits". We wish you to have the sweetest fruits.

Literature

Mosteller, F. a.o. *Probability. A First Course.* Reading, 1961.

Feller, W. *An Introduction to the Theory of Probability and Its Applications.* 3rd ed. Vol. 1, Wiley, New York, 1968.

Feller, W. *An Introduction to the Theory of Probability and Its Applications.* Vol. 2, Wiley, New York, 1966.

David, F. N. *Games, Gods, and Gambling. The Origin and History of Probability and Statistical Ideas from the Earliest Times to the Newtonian Era.* Hafner, New York, 1962.

Parzen, E. *Modern Probability Theory and Its Applications.* Wiley, New York, 1960.

Gnedenko, B. *The Theory of Probability,* 3rd ed. Mir Publishers, Moscow, 1979.

Seber, G. A. F. *Elementary Statistics.* Wiley, Sydney, 1974.

Alder, H. L., Roessler, E. B. *Introduction to Probability and Statistics,* 6th ed. W. H. Freeman and Co., San Francisco, 1975.

To the Reader

Mir Publishers would be grateful for your comments on the content, translation and design of this book. We would also be pleased to receive any other suggestions you may wish to make.

Our address is:

Mir Publishers
2 Pervy Rizhsky Pereulok
I-110, GSP, Moscow, 129820
USSR

Printed in the Union of Soviet Socialist Republics

The author of this booklet describes in popular language how probability theory was developed and found wide application in all fields of modern science.
This book can be considered as an introduction towards a more thorough study of probability theory and is intended for a wide circle of readers.

Mir Publishers Moscow